Winter 1-15-2021

Understanding and Exploiting Protein Allostery and Dynamics Using Molecular Simulations

Sukrit Singh
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS

Division of Biology & Biomedical Sciences
Computational and Molecular Biophysics Program

Dissertation Examination Committee:
Gregory R. Bowman, Chair
Kendall J. Blumer
Rohit V. Pappu
Linda J. Pike
Jay W. Ponder
Janice L. Robertson

Understanding and Exploiting Protein Allostery and Dynamics Using Molecular Simulations

by
Sukrit Singh

A dissertation presented to
The Graduate School
of Washington University
in partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

January 2021
St. Louis, Missouri

Contents

List of Figures

List of Tables

Acknowledgments

When I was younger I somehow found myself constantly going back and forth trying to figure out my favorite subject in science class: biology, chemistry, or physics. This dissertation is the culmination of that debate, born out of a desire to integrate all my interests, and is a dream come true.

This thesis would not have been possible without the help of many mentors, family members, friends, and colleagues. There are too many people to thank; if you are reading this, know that I am grateful for your support and contributions.

Primary thanks go to my mentor Gregory Bowman for pushing me to be the best scientist I can be. Greg gave me the freedom to explore the boundaries of science and the mentoring to grow both as a person and as a scientist. It is fun to reminisce on how we first met when the lab was still covered in cardboard, and the wild journey we've been on since. My present and future success is in large part because of him.

Science can feel like a solo endeavor, but I have been lucky enough to be surrounded by lab members, both past and present, that made me feel like part of a team. I am grateful for the time I got to spend with Xianqiang (Leos) Sun, who I worked closely with on the G protein projects and learned much about computational biophysics from. I also want to thank Katie Hart and Chris Ho, whose advice and guidance during their time in the lab was invaluable. A special thanks goes to Thomas Frederick, for being both an incredible colleague and a wonderful friend. Spending time after hours reading papers or arguing about data was a perfect blend

of fun and insightful. It has also been a joy working with Neha Vithani on subsequent G protein projects. Neha has been a fount of positivity and support for me, providing both personal and scientific advice on all occasions. I am forever grateful for her patience with me pinging her with a tidal wave of questions. Working alongside friends like Matthew Cruz and Catie Knoverek has been an incredibly rewarding experience; Their unconditional love and support kept me going when I struggled and getting to spend time with them, both in the lab and outside, has brought me nothing but joy. It has been a pleasure to grow and learn alongside Maxwell Zimmerman, who joined the lab at the same time I did, and so became my "brother" in the lab. I am grateful to Justin Porter and Mickey Ward for all our wonderful discussions that would range from science to the news to economic systems. I enjoyed spending time with newer members of the lab Artur Meller, Upasana Mallimadugula, Jonathan Borowsky, and Catherine Kuhn, who are doing incredible work and have taught me much in a short time. Lastly, I wish to thank my two "Bay buddies" from when I first started in the lab, Carrie Sibbald and subsequently Katie Moeder. Every conversation with either of you was an enriching experience and made me a better scientist (and arguably a better person).

I want to thank my thesis committee (Linda Pike, Ken Blumer, Jay Ponder, Rohit Pappu, and Janice Robertson) for their mentorship and advice throughout my graduate career. They, along with other faculty members in my department, served as sources of mentorship and insight into a life in science and research. I particularly want to thank Ken Blumer and Jay Ponder, who over multiple conversations gave me tons of advice about an academic career and my career trajectory. Jay, along with Garland Marshall, has been a long-time supporter of mine since my undergraduate years, and I will forever be grateful for his support and mentorship.

I am grateful to have been a part of the Folding@home (F@h) consortium. Getting to discuss science and F@h logistics with PIs and fellow F@h scientists has been an incredibly enriching experience. I learned so much about managing an organization like F@h thanks to Anton Thynell, and about engineering a network of this scale thanks to Joseph Coffland. A big shout-out goes to all the testers, citizen-scientists, and volunteers that have been a part of Fold-

ing@home. Without you, this thesis would literally not have been possible. Discussing science and troubleshooting simulations and projects with testers on Slack and our forums made me feel like they were a part of the lab.

This thesis was possible thanks to support, financial and otherwise, from a variety of sources. I want to thank the National Institutes of Health, the Division of Biological and Biomedical Sciences (DBBS), and Millipore-Sigma for providing much of my graduate funding. Also a quick shoutout to NVIDIA for providing a GTX Titan X to the lab, which was useful for preliminary simulations. I also want to thank the Biochemistry and Molecular Biophysics (BMB) department for the opportunities to be involved in department events through my time on the Student Liaison Committee, and for all the free coffee (that directly led to this thesis). A **big** thank-you to the admin folks in the BMB department, the DBBS program, and the OISS office. They helped my graduate school experience go smoothly and without any major hitches. They have been nothing but helpful and provided resources at every turn.

Having a support network is critical to surviving graduate school, and I would not have been able to complete this thesis without the support of my friends. To Alex Bernstein, Daniel Deibler, Scott Haber, Jack Reidy, and Sid Ravishankar: Thank you for all the years of listening to me talk about science, your unending support when I struggled, and all the endless fun we've had over the years. Thanks to my St. Louis D&D group (Hannah, Jeffrey, Shawn, Anthony, Charley, and River) for all the fun times over the years. Shout out to Becky Ye for all the fulfilling conversations, and to my friends Jake Lyonfields and Ashley Kuykendall for their support and advocacy; I am a better person because of all that you opened my eyes to. Lastly, a big shout-out goes to friends and colleagues in my graduate program (Tyson, Robb, Kacey, Jim, Josh, and Jessey) and beyond (Joseph H., Matt H., Rafal W.) for the years of great conversations, fun discussions, and incredible memories.

I want to give a thank you to the teachers I have had from Singapore American School, American Embassy School (New Delhi), and my undergraduate years at Washington University. I would not be here without the education and direction you provided me. Thank you to Ms.

Jain, Ms. Sosa, Mr. Brakenhoff, Mr. Ortiz, and many others who taught me through the years.

I am grateful to my mom and dad for literally making this thesis possible, and for showing me a world of different cultures and opening my mind to new perspectives. A big shout-out goes to my cousins Kshitij (Conny) and Kapun, who have been nothing but loving and supportive since the day I was born. Final thanks go to my partner Sophia Fox-Dichter for her unending love and support. This thesis is as much her effort as it is mine.

Bonus thanks go to Coco the dog. You have put more into this thesis than you will ever comprehend. You are a good girl.

Sukrit Singh

Washington University in St. Louis

January 2021

Dedicated to all the immigrants out there. You're getting the job done.

ABSTRACT OF THE DISSERTATION

Understanding and exploiting protein allostery and dynamics using molecular simulations

by

Sukrit Singh

Doctor of Philosophy in Computational and Molecular Biophysics

Washington University in St. Louis, 2019

Professor Gregory R. Bowman, Chair

Protein conformational landscapes contain much of the functionally relevant information that is useful for understanding biological processes at the chemical scale. Understanding and mapping out these conformational landscapes can provide valuable insight into protein behaviors and biological phenomena, and has relevance to the process of therapeutic design.

While structural biology methods have been transformative in studying protein dynamics, they are limited by technical limitations and have inherent resolution limits. Molecular dynamics (MD) simulations are a powerful tool for exploring conformational landscapes, and provide atomic-scale information that is useful in understanding protein behaviors. With recent advances in generating datasets of large timescale simulations (using Folding@home) and powerful methods to interpret conformational landscapes such as Markov State Models (MSMs), it is now possible to study complex biological phenomena and long-timescale processes. However, inferring communication between residues across long distances, referred to as allosteric communication, remains a challenge.

Allostery is a ubiquitious biological phenomena by which two distant regions of a protein are coupled to one another over large distances. Allosteric coupling is the mechanism through which events in one region (such as ligand binding) alter the conformation or dynamics of another region (ie. large conformational domain motions). For example, allostery plays a critical role in cellular signaling, such as in the transfer of a signal from outside the cell to cytosolic proteins for generating a cellular response.

While many methods have made tremendous progress in inferring and measuring allosteric communication using structures or molecular simulations, they rely on a structural view of allostery and do not account for the role of conformational entropy. Furthermore, it remains a challenge to interpret allosteric coupling in large, complex biomolecules relevant to physiology and disease.

In this thesis, I present a method to measure the Correlation of All Rotameric and Dynamical States (CARDS) which is used to construct and interpret allosteric networks in biological systems. CARDS allows us to infer allostery both via concerted changes in protein structure and in correlated changes in conformational entropy (dynamic allostery). CARDS does so by parsing trajectories into dynamical states which reflect whether a residue is locally ordered (ie. stable in a single rotameric basin) or disordered (ie. rapidly hopping between rotamers).

Here I explain the CARDS methodology (chapter 2) and demonstrate applications to a variety of disease-relevant systems. In particular, I apply CARDS and other sophisticated computational methods to understand the process of G protein activation (chapter 3), a protein whose mutations are linked to cancers such as uveal melanoma. I further demonstrate the utility of CARDS in the study a potentially druggable pocket in the ebolavirus protein VP35 (chapter 4). The analyses and models constructed in this work are supported by experimental testing. Lastly, I demonstrate how integrating MD with experiments, sometimes with the help of citizen-scientists around the world, can provide unique insight into biological systems and identify potentially useful targets. In particular, I highlight our recent effort converting Folding@home into an exascale computer platform to hunt for potentially druggable pockets in the proteome of SARS-CoV-2 (chapter 7) (the cause of the COVID19 pandemic).

Chapter 1

Introduction

1.1 Protein conformational landscapes encode functional information.

Proteins are the machines that power cellular function and life. They allow us to see, smell, think, and carry out many of the basic functions required for us to live. However, when they malfunction or misbehave (usually through mutation), they can also result in diseases like cancer or heart disease among others. Proteins are also utilized by viruses and bacteria to infect host cells, replicate, or even break down the drugs we use to stop them.

Understanding protein behaviors relevant to health and disease depends on being able to model them in atomic detail. Atomistic scale motions allows us to infer things like mechanisms, thermodynamic profiles, and kinetic rates. This level of detail can provide predictive models and explanations for why certain mutations may cause disease, which can be useful for targeting proteins using drug design methods. Furthermore, knowledge of chemical interactions allows us to design chemical groups against pockets that would jam them open [4]. Atomic-scale knowledge may even provide guiding principles upon which proteins can be designed to

perform novel functions, which has implications for therapeutic design and industrial applications [5].

Structural biology methods have been transformative in allowing us to learn about the structure of proteins and their behaviors. The first view of protein structures was done using X-Ray crystallography [6]. However, static structures do not provide a complete picture of protein function. These static structures may not be able to provide information about a proteins stability [7], ligand affinities or specificities [8], or how different mutations could impact function [9]. Indeed, it has been often observed that crystal structures of the same protein families are too similar to explain the difference in their measured physiological parameters [10].

As the atoms of a protein move around relative to one another, a protein is able to shift between an enormous number of different structures. Even small proteins of $<$100 amino acids have $\sim$200 rotatable bonds along their backbone, granting access to more than 10^{60} backbone conformations [11]. Many of these structures however are never accessed by the protein in physiologically relevant timescales. Of the fraction that are accessed however, some of them may have relevance to a protein's mechanism and biophysical behavior. Each of these structures that a protein may shape-shift into has an associated energy that characterizes intra-protein and protein-environment interactions. Given that the probability of a protein adopting any one structure is proportional to the exponential of that structures energy, we are able to characterize how likely a protein is to adopt some states over others. The phase space of a protein's energies (or probabilities) and their corresponding structures are often referred to as an "energy landscape", with most likely states (such as those observed by crystallography) are named "ground" states due to being energy minima.

From these ground states, a protein can also transition into less likely "excited" states, some of which may contain key functional information. These states can be characterized using a plethora of methods. Nuclear Magnetic Resonance (NMR) and Hydrogen-Deuterium Exchange (HDX) have provided unique functional insight into the conformational heterogeneity

a protein can have [12, 13]. Work on DHFR has been critical in understanding the complete catalytic cycle and the residue level configurations (and dynamics) [14]. NMR experiments on DHFR that was arrested in one stage of its catalytic cycle provided evidence that DHFR was also adopting stages in the latter part of its cycle [14]. Mutational experiments also put forth a correlation between dynamics and catalytic ability [13–15]. Powerful combinations of NMR, crystallography, and computer simulations have rationalized cofactor- and mutational-effects on kinases [16,17]. With advances in structural methods, it is even possible to resolve structural information about excited states [14,15]. Other electron paramagnetic resonance (EPR) experiments have granted unique insight into the conformational distributions of proteins [18]. NMR also has the additional power to measure the degree of conformational entropy of residues [19]. Recently, Cryo-EM structures of even large complexes have revealed the degree of conformational heterogeneity in physiologically relevant systems [20–23]. However, it is important to note that each of these methods have tradeoffs due to resolution limits, labelling strategies that may perturb the system, or technical limitations (system size, material requirements, etc.). Altogether, this large body of work studying excited states of folded proteins suggests that the equilibrium motions of a protein may encode all of its functionally relevant states.

Molecular dynamics (MD) simulations have the potential to provide atomistic detail to explain complex biological processes. These simulations compute the movements of atoms over time by integrating Newton's laws of motion over each atom. Thus, MD acts as a "computational microscope", allowing us to observe the different conformations a protein adopts [24, 25]. A perfect simulation would completely describe a protein's thermodynamic and kinetic behaviors at equilibrium. However, there are major limitations: (i) The accuracy of atomic parameters (aka "force fields") that are used to describe atomic interaction (A topic that is discussed extensively elsewhere [26]). (ii) Simulations take femtosecond-sized timesteps, making it expensive to gather data at biologically relevant timescales (microsecond to milliseconds). (iii) Interpreting large datasets with thousands of unique structures to generate biologically meaningful predictions remains a daunting task.

1.2 The Folding@home platform allows access to protein motions at biologically relevant timescales.

1.2.1 Folding@home distributes simulations across thousands of computers at once.

A myriad of methods have been developed to improve sampling of protein motions out to biologically relevant timescales [27–30]. However, many perturb the thermodynamics or kinetics of a system, in turn biasing the predictions these simulations make. The advent of GPUs provided access to longer timescales, but they are bounded by an upper limit of parallelism in computing architectures [31–34]. Recently, novel adaptive sampling techniques have allowed for the mapping of slower motions [35] but may require system-specific knowledge or an order parameter of relevance. Specialized hardware has also been developed [36] that allowed computational biophysicsts to study processes at unprecedented timescales. However, utilizing and maintaining this kind of specialized hardware can be a costly endeavor.

To sample unbiased simulations at biologically relevant timescales using commodity hardware, the Folding@home platform was developed in 2000 [37]. Folding@home, headquartered at the Bowman lab at WUSTL, is a distributed computing network that runs MD simulations on donated computing power thanks to thousands of citizen-scientists who download the app onto their hardware. Folding@home runs trajectories as small "work units" that are smaller simulations on the order of nanoseconds. The starting file for a work-unit is generated server-side, which is sent to a client computer somewhere in the world (Fig. 1.1). The client then runs the work-unit and returns it. This is used to generate the subsequent work-unit (representing the next chunk of time in a simulation). Afterwards, work units are stitched together to generate a single trajectory. Folding@home allows for the generation of a datasets containing hundreds of trajectories that, in aggregate, capture a large amount of a protein's energy landscape. These datasets are so large that interpreting them presents a unique "big data" challenge, and requires

the development of new methods and software [38, 39].

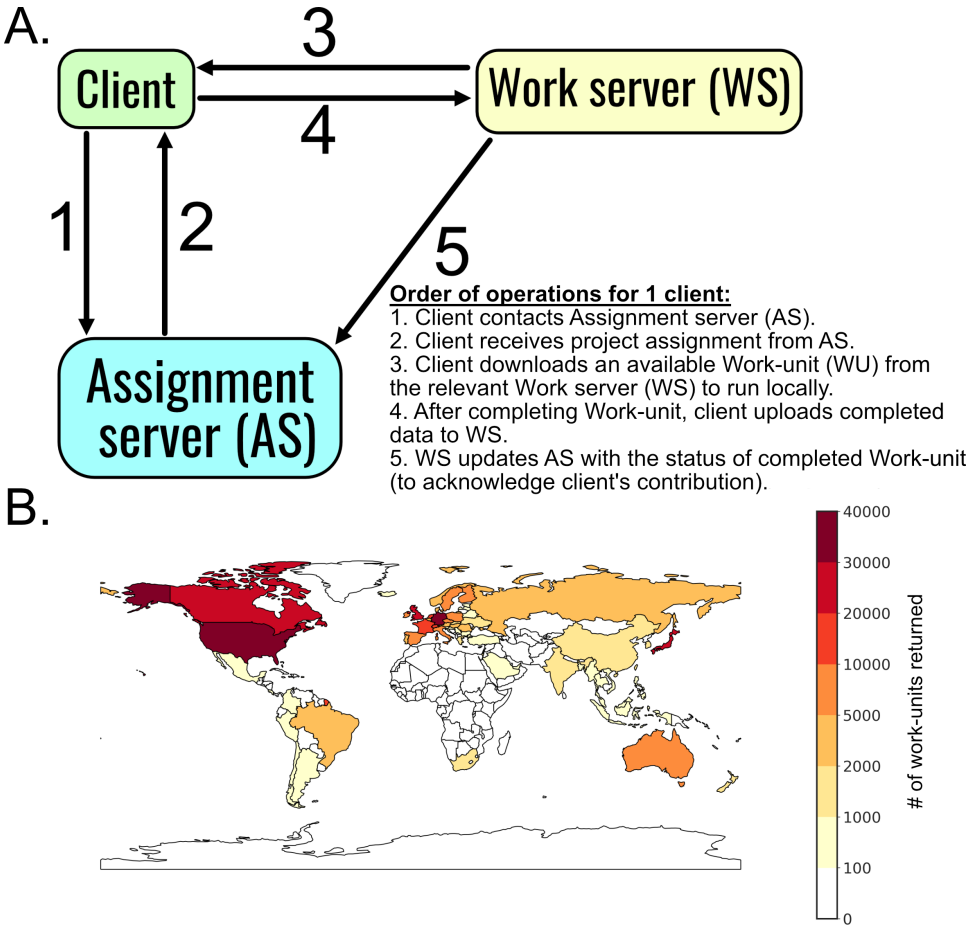


Figure 1.1: Folding@home distributes work units worldwide to run simulations **A.** Workflow schematic detailing the steps a citizen-scientist's computer ("client", green) takes to receive, run, and return a work-unit successfully by communicating with the Assignment Server (blue) and the Work server (yellow). **B.** Heatmap of completed work-units returned from each country in the world in a representative 48-hour period. The number of returned work units is indicated by the color scale (right).

1.2.2 Markov State Models allow for the construction of unified models from large simulation datasets.

Markov state models (MSMs) are network representations of a protein's free-energy landscape, providing map representations of protein conformational space with thermodynamic and kinetic properties taken from equilibrium simulations (Fig. 1.2). Rather than depending on a

single long simulation that explores multiple states sequentially, MSMs are capable of stitching together multiple short trajectories into a single unified landscape. Thus MSMs capture slow events, and their intermediates, far beyond the reach of any individual simulation. Thanks to distributed networks like Folding@home, gathering large numbers of short trajectories is tractable in a reasonable time-frame. Many reviews have been dedicated to providing accessible and in-depth explanations of MSM technology [38, 40, 41].

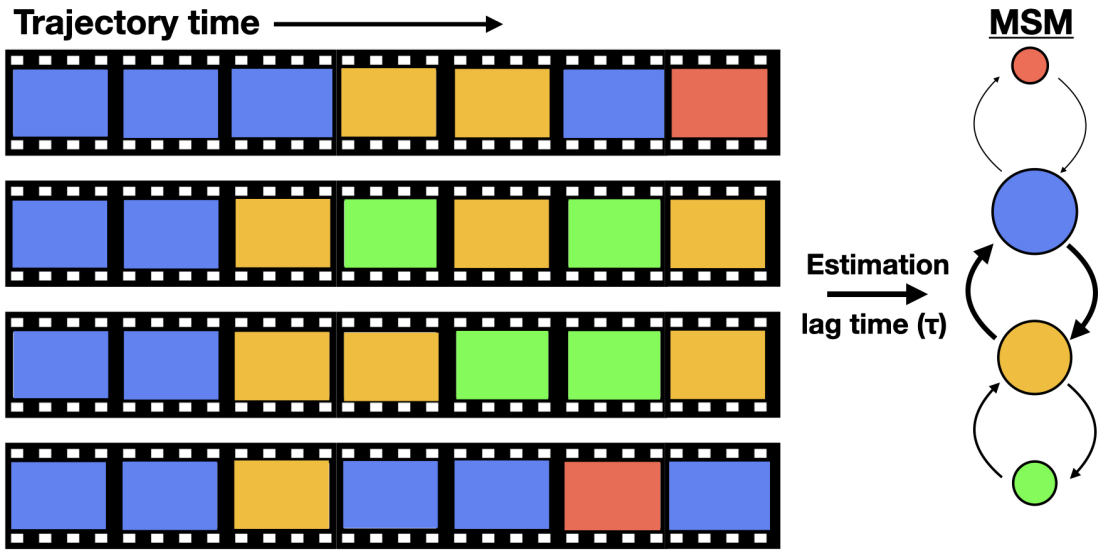


Figure 1.2: Diagram highlighting the conversion of a simulation dataset into a Markov State Model (MSM). Simulation trajectories (left) are parsed into discrete states on a per-frame basis (red, blue, green, yellow), and estimation methods are used to convert the trajectories into a MSM (right). The spheres radii is proportional to the population of that state, while arrow thickness denotes transition probability.

MSMs require two components to describe the dynamics of a biomolecular system: (i): A discretization of a high dimensional state space into n conformational states, and (ii) A model of the stochastic transitions between each states, represented as a Transition Matrix **P**. The probability of transitioning from state i to state j (P_{ij}) is:

$$P_{ij}(\tau) = Prob(x_{t+\tau} \in S_j | x_t \in S_i) \tag{1.1}$$

where τ is a lag-time parameter across which transitions between states are observed, and S_i is state i and S_j refers to state j. These transition matrices $\mathbf{P}$ give rise to a stationary distribution π as a result of the eigenvalue equation:

$$\pi^T \mathbf{P} = \pi^T \tag{1.2}$$

where T represents a single time-step. In an MSM, the stationary distribution π represents the equilibrium probabilities for each state. Therefore, assuming sufficient statistics are collected to observe transitions between states, MSMs are able to appropriately capture the equilibrium thermodynamic and kinetic behaviors of a system. It is important to note that the underlying assumption behind MSM construction is that the discretized dynamics of biomolecules is memoryless (aka Markovian). That is, the probability of a transition from state i to state j at time t is only dependent on state i and not any of the previously visited states.

For any constructed MSM, the Markovian assumption is tested by looking at implied timescales plots of a Markov State Model. That is, for a given state decomposition, the molecular relaxation timescales for eigenvalues λ and eigenvectors r_i are computed for a series of lag-times τ:

$$t_i = -\frac{\tau}{\ln |\lambda_i(\tau)|} \tag{1.3}$$

These timescales t_i are plotted for a series of lag-times τ (C.8). If the timescales remain relatively unchanged for a series of τ values and higher, then the models constructed for those values of τ can be considered Markovian. The values of P_{ij} can then be estimated using a variety of estimation methods [40, 42].

A critical component of high-quality MSM construction is the features selected in the discretization of state space. It is important to balance statistical error against systematic bias when choosing, as narrow features ranges may result in poor statistics, while broadly-ranged features may have large systematic errors due to convolution of multiple motions into a single state. Choice of appropriate feature can challenging due to the sheer size of simulation datasets

and the system-specific knowledge often need to identify appropriate features.

While methods such as PCA might provide some degree of feature-reduction, a major breakthrough in simulation featurization was the use of time-lagged independent component analysis (tICA) [43], which transforms input coordinates to identify the rarest (which are assumed to be the slowest) motions. For protein folding simulations, this provides excellent dimensionality reduction, since the slowest coordinate is often the rarest and most valuable [44, 45]. However, in simulations of folded proteins the rarest motions observed may be less interesting or artifacts of sampling due to the simulation size. Much success has previously been seen using geometric features, such as cartesian coordinates or dihedral angles [46–48]. Other features such as solvent accessible surface area (SASA) or ligand-residence-time have also been useful for construction of predictive MSMs [48, 49].

Once appropriate features are selected, there are a myriad of approaches to cluster them into discrete states. One major method is a hybrid k-centers/k-medoids approach [50, 51]. In brief, for every pair of features a distance metric, such as the Root Mean Square Deviation (RMSD) is computed. The k-centers algorithm then (i) chooses an initial cluster center either as a predetermined point or randomly, and then all points are assigned to this initial cluster. (ii) The distance between every point and its assigned cluster center is then computed. (iii) The point with the largest distance to the assigned cluster center is then labelled as a new cluster center. (iv) The distances between all points and all cluster centers are recalculated, and points are reassigned to their closest cluster center based on these new labels. Steps $(ii - iv)$ are repeated until the maximum distance from any point to its assigned cluster center goes below a specified threshold, or a maximum number of cluster centers is reached.

To refine the cluster center assignments, and ensure that the "center" of each cluster is truly equidistant from all assigned points, a k-medoids algorithm known as Partitioning around Medoids (PAM) [51] is used. PAM proceeds by iterating through each cluster and choosing a new center from one of the currently assigned points. All points across all states are then

reassigned based on this new proposed cluster center. From this proposed center, a "cost" is calculated (i.e. the sum of distances from each point to their respective center), and the proposed center placement is accepted if the cost is minimized. Once states are discretized, clustered, assigned across a simulation trajectory, the transition probability matrix is estimated. As mentioned above, the matrix of transitions between states is counted based on some lag time τ.

One simple approach to estimate transitions is to count the number of transitions from states i to j between all states, and divide by the number of states i observed. To maintain ergodicity, some methods only estimate over the largest connected subset of states [52] or using maximum likelihood estimators that respect detailed balance [53]. One method to estimate transition probabilities is to to average the count matrix with the transpose of itself

$$C_{ij}^{transpose} = \frac{C_{ij} + Cji}{2} \tag{1.4}$$

where C_{ij} is the observed number of transitions from state i to state j, and C_{ji} represents the number of transitions in the reverse direction. Subsequent row normalization is used to calculate the equilibrium probabilities:

$$\pi_i = \frac{\sum_j C_{ij}^{transpose}}{\sum_{k,j} C_{k,j}^{transpose}} \tag{1.5}$$

where $C_{ij}^{transpose}$ is the averaged number of transitions between states i and state j, and $C_{k,j}^{transpose}$ is the number of transitions between states k and j. Recent success has been observed by simply adding a pseudocount $\tilde{C}$ to serve as an estimate of the system in absence of data [42, 54]. This pseudocount is computed as a single observed transition that is divided up across all states

$$\tilde{C} = \frac{1}{N} \tag{1.6}$$

where N is the number of states.

1.2.3 The scalable power of Folding@home has generated insights into protein behaviors.

There are many success stories that have come out of the usage of the Folding@home platform such as the observation of a millisecond-timescale folding event in 2010 [46]. Markov state models have been shown to quantitatively agree with experimental measurements [40,55]. Particularly, there is excellent agreement between microsecond-scale simulations and the properties of systems measured with NMR and room-temperature crystallography [56, 57]. Furthermore, simulations have been able to interpret the impact of mutations on diseases such as phenylketonuria [58] and characterize the landscapes of a myriad of targets [59]. Folding@home and MSMs have allowed for the assessment of families of protein homologs [10].

Folding@home has also made significant progress in developing and supplementing experimental work with predictive models. For example, Folding@home data was used to identify novel cryptic pockets in TEM-1 β-lactmase [49], the enzyme most directly involved in antibiotic breakdown and microbial resistance. Indeed, accounting for the dynamics within the active site of TEM-1 β-lactamase substantially improved the predictive ability of modern virtual screening technologies [9]. Similar approaches have yielded valuable insights into the pH dependence of protein-protein interactions [60], and other biological phenomena.

1.3 Allosteric communication is critical for protein function, but difficult to infer.

1.3.1 Allosteric communication is universal and critical for biological function.

One cellular phenomena existing at long-timescales is communication between distant structural elements of a protein. This behavior, referred to as **allostery** [61], was first recognized in hemoglobin [62] where the binding of oxygen to a single subunit increases the oxygen affinity within the other three subunits. Since then, the importance of allostery has been recognized in a myriad of cellular functions, such as transcription factors [63, 64] or cellular signaling [65].

A well-studied protein (and drug target) with allosteric behavior is the G protein coupled receptor, which transmits information from outside the cell to inside based on a stimulus. This stimulus can be anything from ligand binding [66] to membrane deformations [67]. Structural methods revealed rearrangements of transmembrane helices that convert the GPCR into an "active" form [68]. However, recent data of different GPCR-Gα complexes have highlighted the conformational heterogeneity in the allostery and activation of GPCRs [20–23].

The idea that a protein's conformational landscape can impact its allosteric behavior leads one to speculate if all proteins have some degree of allosteric coupling [69]. Indeed, the ubiquity of allostery has been acknowledged in studies of natural and directed evolution [70, 71], where mutations distant from the active site can impact measured properties. Given the potentially universal nature of allostery, it is worth speculating if mutations and ligands work by tapping into existing allosteric networks to modulate the distributions of a protein's of structures and dynamics [41]. Thus, understanding allosteric coupling in proteins could present new opportunities for modulating biological processes, designing therapeutics, or even designing new proteins.

Mounting evidence highlights the value in leveraging allostery to modulate protein function. There are a multitude of biological systems where allostery is leveraged to inhibit or activate protein-protein interactions, and so it may be possible to identify small molecules that could achieve the same objective of modulating protein behaviors. An allosteric drug that could modulate protein behaviors could play a huge role in restoring lost functions or reducing over-active protein behaviors. However, modern drug-design methods often require the presence of a cavity (or "pocket") to successfully design small molecule hits. Many surfaces involved in protein-protein interactions are often too flat for a small molecule to bind tightly [72], and targeting known ligand binding sites of critical signaling proteins like GPCRs creates the risk of off-target effects. Identifying distant pockets that are not as conserved between homologs could be a means to achieve specificity [73].

Hidden allosteric sites known as 'cryptic pockets' could be promising targets for drug design methods. The shape-shifting nature of proteins implies the existence of states that contain new pockets that are not observed in existing experimental structures. The hidden pockets may also be cryptic allosteric sites that are connected to key functional sites via the underlying allosteric network of a protein [74]. Successful methods have emerged to identify novel allosteric sites, some of which have been verified by experiments [49]. Indeed, the value of cryptic pockets has been supported by the discovery of small molecule inhibitors that are shown to bind an allosteric pocket and modulate a protein's function [9, 75–77]. Computer simulations provide promising avenues to hunt and target cryptic pockets, an effort which has yielded promising results [78, 79], but it remains a challenge to apply these approaches to a wide variety of systems. Furthermore, the discovery of a distant pocket in a protein does not imply it is "useful" as a drug target, because it is difficult to measure the degree of coupling between the cryptic pocket and a protein's functional regions (like active site). Understanding the communication between a cryptic pocket to functional regions of a protein could further supplement drug design strategies. However, obtaining a complete picture of the allosteric network of a protein is often difficult.

1.3.2 Inferring allosteric communication in proteins remains a non-trivial task.

Methods for inferring allostery typically rely on observing concerted structural changes. A system with two distant sites may jump between alternative configurations in some coupled fashion. That is, the structure of site A may be coupled to the configuration of site B, and vice versa. This extreme example of conformational selection could be inferred by comparing structures of proteins before and after some perturbation to one of the sites is introduced (such as ligand binding). Indeed, crystallography and HDX methods have proved useful in revealing residues involved in allosteric networks of TIM-barrels and their catalytic domains [80]. Multiple NMR methods have also proven useful in studying the nature of allosteric communication between proteins [81].

Likewise, computational methods measure concerted structural changes using a variety of metrics on a myriad protein features. Some methods utilize sequence coevolution to group proteins regions into "sectors" that are coupled to one another [82]. Recent work has highlighted that molecular simulations can capture atomistically detailed pictures of allosteric coupling between sites. The underlying assumption that a proteins functional states are encoded in the equilibrium simulations implies that observing correlated motions in MD simulations would be representative of the degree of coupling between residues. Indeed, a number of algorithms use a myriad of features and metrics to quantify coupling [83,84]. Some features used could be the backbone Cα atoms of proteins, and measuring the degree of covariance in pairs of Cα atoms [85]. Other methods utilize mutual information methods on dihedral angles to quantify how much better one residue's dihedral angle predicts the dihedral angle of another residue [86]. However, there has been growing recognition that allostery via concerted structural changes is not the only mechanism through which two sites may be coupled.

In recent years the role of conformational entropy in allosteric communication has been increasingly acknowledged. The importance of conformational entropy was first described theoreti-

cally in 1984 by Cooper and Dryden [87]. Since then experimental evidence for this "dynamical allostery" has grown. Particularly, NMR data demonstrated two sites on a transcription factor were coupled with no discernible structural changes [88]. Furthermore, intrinsically disordered regions can also play a role in allosteric coupling, as normally ordered regions of a protein may transition locally into higher-entropy excited states that rapidly hop between multiple thermodynamic minima. This flattening of the effective free energy surface for a set of residues distinguishes dynamic allostery from the previously discussed mechanism of concerted structural changes. More recently, it has become apparent that to understand a protein's allosteric network, it is important to observe both concerted structural changes and altered conformational entropy [89]. The ability to construct allosteric networks, by measuring both structure and disorder, has the potential to explain the mechanism of coupling in many complex biological process and may present opportunities to identify promising druggable pockets and the role they play in modulating protein function.

1.4 Scope of thesis.

It remains a critical challenge to understand allostery to completely describe biological behavior. The potential power of MD simulations to explain complex biological processes in atomistic detail presents a promising avenue to achieve these goals, but the tools to do so remain limited in scope, and generating simulation datasets that capture slow allosteric processes remains difficult. This thesis describes an approach to understand allosteric communication and the conformational landscape of proteins, and leverages these insights towards understanding fundamental biological phenomena or supplementing drug design efforts.

In this thesis I will describe a method to infer allosteric coupling in MD simulations via both concerted structural changes and conformational entropy. This will be done by measuring the Correlation of All Rotameric and Dynamical States (CARDS) – a novel method presented in **chapter 2**. This algorithm builds upon previous works that infer allostery through structural

changes by capturing allostery through changes in conformational entropy. It parses dihedrals into dynamical states, capturing whether a rotamer is ordered (remaining in a single basin) or disordered (rapidly hopping between basis). We then describe our framework to measure coupling between every pair of residues by computing communication between rotameric states, between dynamical states, as well as cross-correlations. We apply the CARDS methods to a system with known dynamic allostery, the Catabolite Activator Protein (CAP), a transcription factor whose allosteric behavior was previously measured in NMR and ITC studies.

Chapter 3 describes the application of CARDS and other MD/MSM methods to a known allosteric system of importance, the heterotrimeric G proteins. Heterotrimeric G proteins are molecular switches that regulate everything including vision, smell, and neurotransmission. Malfunctions in G protein activation are implicated in cancers such as uveal melanoma. While considerable work has characterized G protein thermodynamics and kinetics, a complete mechanism of activation remains unclear including the allosteric network coupling the receptor and nucleotide binding sites. Here we describe in atomistic detail for the first time a complete mechanism of G protein activation, GDP release, and the conformational and dynamical changes driving this process.

Chapter 4 describes the discovery of a hidden cryptic pocket in the previously 'undruggable' ebolavirus protein VP35. The seeding strategy described in chapter 3 is applied to the RNA-binding VP35 protein. A cryptic pocket is identified from a Folding@home dataset, and CARDS identifies the degree of coupling between the cryptic pocket and residues important for Protein-Protein and Protein-Nucleic-Acid interactions (PPIs and PNIs, respectively). The existence of this cryptic pocket is supported using experiments, and the functional importance of this distant allosteric site is solidified using experimental techniques that observe VP35 inhibition after pocket-open state is stabilized.

Chapters 5 and 6 describe recent efforts utilizing Folding@home to study SARS-CoV-2, the virus behind the COVID19 pandemic. In **chapter 5**, we rapidly generate conformations of

the Nucleocapsid protein folded domains, and integrate them with Monte Carlo simulations and experiments. This study shows that the Nucleocapsid protein is dynamic, disordered, and undergoes Liquid Liquid Phase Separation (LLPS) behavior. Folding@home simulations of the folded domain are used to seed Monte Carlo simulations of the intrinsically disordered regions of the Nucleocapsid protein, obtaining a complete picture of the free energy landscape; a feat which would not have been achieved using a single method. These simulations are integrated with experimental data to describe a model explaining how the N protein may package the genome.

In **Chapter 6**, I describe a recent effort where Folding@home shifted focus to simulating potential drug targets in the proteome of the SARS-CoV-2 virus, the cause of the COVID19 pandemic. Many citizen-scientists around the world rallied together, downloading the Folding@home app and running simulations of almost every possible protein from SARS-CoV-2. This effort generated a historic 0.1 seconds of equilibrium simulation data using Folding@home. Included is a description of how, through generations donations and partnerships, Folding@home surpassed the exascale barrier in computing speed, a feat never before achieved in human history. We utilize this monumental computational power to study how the viral Spike protein uses conformational masking to evade an immune response, and identify cryptic pockets that are not present in existing experimental snapshots. The data generated by Folding@home presents new potential targets for drug design efforts, and new structural and mechanistic insights that may supplement the design of therapeutics.

Chapter 7 demonstrates how MD simulations can be integrated with standard structural biology techniques to explain mechanisms of antibiotic resistance. The cefotaximase enzyme CTXM is responsible for the breakdown of many modern cephalosporins, which can result in microbial resistance and sustained infections. With each new drug generated, such as Ceftazidime (CAZ), CTXM has been shown to accrue mutations that grant it enhanced resistance profiles. This chapter will focus on two mutations, D240G and P167S. Counter-intuitively, these mutations are not additive in their behavior, and the presence of both mutations abrogates

the CAZ resistance profile in CTXM. Combining MD and crystallographic methods, along with biochemical approaches, we describe how these mutations alter the acyl-enzyme complex and modulate a key-region in CTXM known as the Ω-loop. We show that while each mutation uniquely modulates the Ω-loop to better accomdate CAZ, both mutations revert the conformational behavior of the Ω-loop back to its non-resistant state. This study demonstrates the unique ability of combining MD with structural biology methods to better understand fundamental behaviors and biological phenomena.

Lastly, in **chapter 8** I summarize the main advancements presented within this thesis. This chapter further explores the general implications of these findings and how allosteric coupling may be a universal phenomena. I expand by discussing future projects, discussing future promising questions, and speculate on the prospect of further integrating simulations with experiments to better answer fundamental biological questions.

Chapter 2

Quantifying allosteric communication via both concerted structural changes and conformational disorder

The work in this chapter is published in: Singh, S., and Bowman, G.R., Quantifying allosteric communication via both concerted structural changes and conformational disorder with CARDS. Journal of Chemical Theory and Computation. 13:1507-1517, 2017. PMID: 28282132., Copyright 2018 American Chemical Society [90]

2.1 Abstract

Allosteric (i.e. long-range) communication within proteins is crucial for many biological processes, such as the activation of signaling cascades in response to specific stimuli. However, the physical basis for this communication remains unclear. Existing computational methods for identifying allostery focus on the role of concerted structural changes, but recent experimental work demonstrates that disorder is also an important factor. Here, we introduce the Correla-

tion of All Rotameric and Dynamical States (CARDS) framework for quantifying correlations between both the structure and disorder of different regions of a protein. To quantify disorder, we draw inspiration from methods for quantifying "dynamic heterogeneity" from chemical physics to classify segments of a dihedral's time evolution as being in either ordered or disordered regimes. To demonstrate the utility of this approach, we apply CARDS to the Catabolite Activator Protein (CAP), a transcriptional activator that is regulated by Cyclic Adenosine MonoPhosphate (cAMP) binding. We find that CARDS captures allosteric communication between the two cAMP-Binding Domains (CBDs). Importantly, CARDS reveals that this coupling is dominated by disorder-mediated correlations, consistent with NMR experiments that establish allosteric coupling between the CBDs occurs without a concerted structural change. CARDS also recapitulates an enhanced role for disorder in the communication between the DNA-Binding Domains (DBDs) and CBDs in the S62F variant of CAP. Finally, we demonstrate that using CARDS to find communication hotspots identifies regions of CAP that are in allosteric communication without foreknowledge of their identities. Therefore, we expect CARDS to be of great utility for both understanding and predicting allostery.

2.2 Introduction

Despite its fundamental importance, understanding of the physical mechanisms of allosteric communication remains incomplete. For example, significant effort has gone into studying how G-Protein Coupled Receptors (GPCRs) allosterically transmit extracellular signals to intracellular binding partners [91]. However, understanding of this process is still insufficient for the routine design of drugs that allosterically modulate GPCRs [92].

Models of allostery have typically focused on concerted structural changes [93]. For example, the classic induced fit model postulates that ligand binding to one subunit of a protein causes a conformational change in other subunits [94, 95]. The conformational selection model also focuses on structural changes, positing that ligand binding to one subunit stabilizes an alter-

native (but pre-existing) structure of other subunits [96]. Extensive work establishes there is often a role for conformational selection [97], though there is clearly a continuum between extreme versions of induced-fit and conformational selection [98–100]. This conclusion implies allostery can be inferred from proteins' equilibrium fluctuations even in the absence of an allosteric perturbation. Inspired by this implication, numerous methods have been developed to infer allosteric communication from structural fluctuations observed in molecular simulations [83, 85, 86, 101–119].

While concerted structural changes are clearly important for allostery, there is mounting evidence that conformational disorder has an important role to play, and can even lead to allosteric communication in the absence of concerted structural ,transitions [64, 93, 120–125]. The importance of allostery without conformational change first appeared in a model where ligand binding perturbs the entropy of a distant site rather than its preferred structure [87]. Since then, NMR and ITC experiments on Catabolite Activator Protein (CAP) have established allosteric communication without conformational change exists in nature [19, 88, 126]. CAP is a homodimeric transcription factor whose DNA-binding affinity increases upon binding of cAMP to the cAMP-Binding Domains (CBDs) [127, 128]. In wild-type CAP, cAMP binding allosterically induces the DNA-Binding Domains (DBDs) to swivel around the central hinge region into a DNA-binding conformation (Fig. 2.1) [129, 130]. There is also negative coupling between the two CBDs [127, 128, 131]. A combination of NMR and ITC measurements reveal that binding of cAMP to one CBD reduces the cAMP-binding affinity of the second CBD without changing its structure [88, 131]. Additional experiments reveal that binding of cAMP activates the S62F variant of CAP without causing a conformational change in the hinge or DBDs [63, 132].

While the importance of disorder is gaining widespread acceptance, the field lacks systematic methods for identifying allosteric communication in the absence of conformational change. For example, NMR has yet to uncover how these signals are transmitted. COREX/BEST [121, 133, 134], other coarse-grained models [107], and normal modes [135] provide valuable insights but miss essential subtleties, such as important side-chain motions. Using molecular

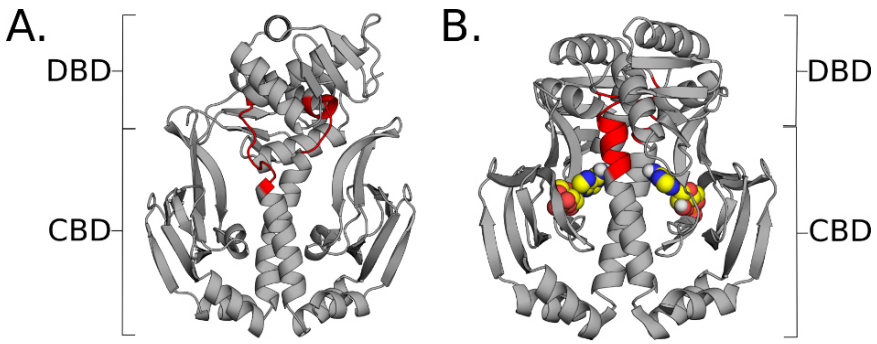


Figure 2.1: Structures of CAP in apo and holo forms. **A.** Structure of apo-CAP, with cAMP-binding domains (CBDs) and DNA-binding domains (DBDs) indicated by brackets. Residues that make up the hinge region are colored in red. **B.** Structure of holo-CAP with cAMP ligands (spheres) bound to the CBD regions. The same hinge residues are colored in red.

dynamics simulations to measure the mutual information between the orientations of different dihedral angles captures the reduction in uncertainty (measured with an entropy metric) about the structure of one dihedral given knowledge of another [86, 108, 109], but does not capture phenomena like changes in the rotameric state of one dihedral increasing the conformational heterogeneity of a distant site. Approaches for inferring allosteric coupling from sequence covariation [82, 136] are agnostic to the mechanism underlying this communication and cannot explain how it occurs. A method for identifying timing correlations has promise for capturing disorder-mediated coupling [89]. For example, application of this approach to side-chain degrees of freedom in CAP successfully identified hotspots for allosteric communication. It also demonstrated that disorder-mediated correlations give rise to long-range communication, while purely structural correlations are limited to short-ranged communication. However, timing correlations do not integrate structural and dynamical correlations into a holistic measure of communication that can capture the continuum of possibilities between purely structural and purely disorder-mediated coupling.

Here, we introduce the CARDS (Correlation of all Rotameric and Dynamical States) methodology for quantifying the roles of both concerted structural changes and conformational disorder. CARDS is based on our observation that a single degree of freedom (e.g. a dihedral angle) can transition between "ordered" regimes, wherein it undergoes small fluctuations within a single

structural state, and "disordered" regimes wherein it undergoes bursts of transitions between different structural states (Fig. 2.2). Similar "dynamic heterogeneity" [137–139] has been observed in the physics of glasses, where it has been shown that a single degree of freedom's local environment can either facilitate dynamics by flattening out the effective free energy surface that degree of freedom experiences or freeze out dynamics [140–142]. CARDS identifies ordered and disordered regimes based on two kinetic signatures: the average time a degree of freedom persists within a structural state and the typical timescale for transitions between structural states. For many dihedrals, we observe that the typical time that elapses between structural transitions, which is dominated by segments of a trajectory in disordered regimes, is orders of magnitude smaller than the typical persistence time in a state. Based on these kinetic signatures, CARDS assigns segments of trajectories to dynamical states (i.e. ordered and disordered regimes). CARDS then quantifies correlations between the structural and dynamical states of different dihedrals. Specifically, we employ the mutual information to assess how much better one can predict the structural and dynamical states of one dihedral given knowledge of the structural and dynamical states of another dihedral. To demonstrate the utility of CARDS, we assess whether it can identify allosteric communication in the absence of concerted structural changes observed in CAP.

2.3 Theory and methods

CARDS captures all forms of correlated fluctuations, including concerted structural changes, correlations between the conformational disorder of different degrees of freedom, and correlations between the structure of one degree of freedom and the conformational disorder of another. As in other recent work, we focus on dihedral angles, as they are natural degrees of freedom for describing protein structure and dynamics and are easily decomposed into a small number of rotameric states [86, 109].

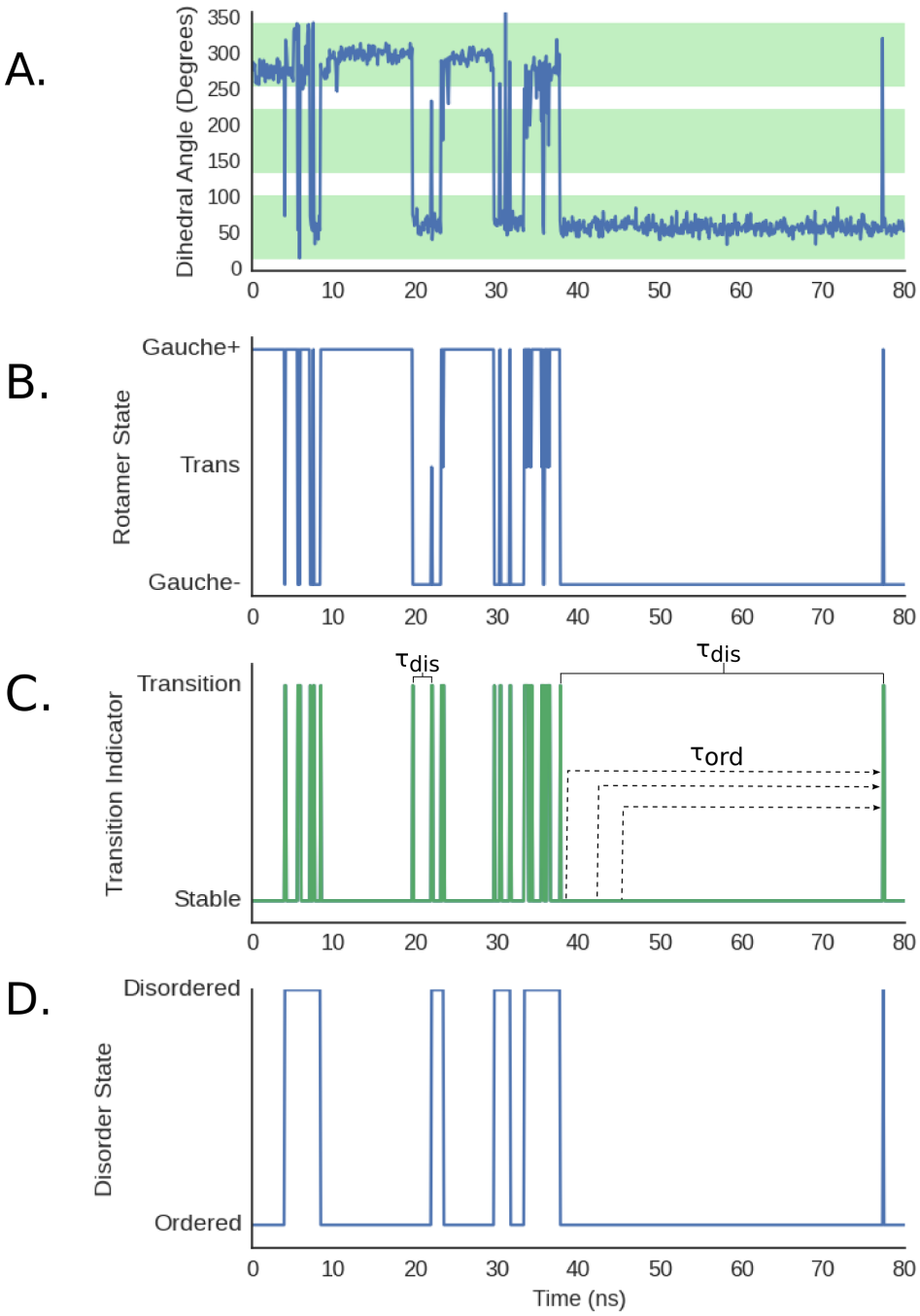


Figure 2.2: Workflow for identifying ordered and disordered regimes. **A.** An example of the time-evolution of a χ_1 angle. The core of each rotameric state is shaded green and the buffer zones between them are white. **B.** Assignment of each snapshot in the trajectory to rotameric states. **C.** Transition indicator function identifying where the dihedral changes states. Examples of ordered (τ_{ord}) and disordered (τ_{dis}) times are labeled. **D.** Assignment of each snapshot in the trajectory to dynamical states (e.g. ordered and disordered regimes).

2.3.1 Molecular Dynamics simulations

We ran three 500 ns simulations with our previously published protocol [9] (see SI for details). Briefly, we placed PDB ID 4N9H [130] in a dodecahedron box and solvated it with TIP3P explicit solvent [143] extending 1 nm beyond the protein in any dimension. We used PyMOL [144] to mutate Ser62 to Phe to create a starting structure for the S62F variant. For each system, we ran simulations with the Gromacs software package [145] using the Amber03 force field [146]. This combination of software and parameters was selected because it has proven reliable in our past studies of both protein folding [55] and structural fluctuations within folded proteins [9,56,147]. As described in the Results section, our microsecond timescale simulations are sufficient to recapitulate much of what is known from experiments about allostery in CAP. The correlation coefficient between the couplings obtained from any individual simulation and the combination of three simulations is 0.64 ± 0.02, suggesting that hundreds of nanoseconds of simulation may be sufficient to discover gross patterns of communication but are insufficient to obtain converged results. The correlation coefficient between the coupling obtained from any pair of simulations and the combination of three simulations is 0.79 ± 0.02, suggesting that 1-1.5 microseconds of simulation give more reproducible results. Taken together with our past work, we conclude that a few microseconds of simulation are adequate to demonstrate the utility of our new method and, likely, to gain valuable insights into many systems.

2.3.2 Assignment of dihedrals to rotameric states

Dihedral angles were calculated with MDTraj [148] and assigned to discrete rotameric states (e.g. gauche+, gauche-, and trans for most χ angles and cis or trans states for backbone dihedrals). Transition-Based Assignment (TBA) is used to distinguish lasting transitions from transient fluctuations [149–152]. TBA prevents over-counting of transitions (e.g. due to fluctuations at a barrier peak where a simulation repeatedly crosses a hard cutoff between rotameric states) by defining a core region within each rotameric state and buffer zones between them. A

dihedral is only considered to have changed rotameric states if it transitions from the core of one state to the core of another state, passing completely through the buffer zone between cores (Fig. 2.2A and 2.2B). A dihedral that starts in one core, enters a buffer zone, and then returns to its initial core is said to have remained in the initial rotameric state. We define the core of a rotameric state as a region of width W centered between the boundaries between rotamers. We present results using a core width of 90°, but our results are robust to variations in the core width ranging from 60° to 90° (Fig. A.1A)

2.3.3 Assignment of snapshots to dynamical states

CARDS assigns snapshots to ordered and disordered regimes based on two variables that describe the dynamics of the trajectory: the mean ordered time (τ_{ord}) and mean disordered time (τ_{dis}). An ordered time (τ_{ord}) is the time from any time-point to the next time where a transition occurs and a disordered time (τ_{dis}) is the time between two consecutive transitions (Fig. 2.2C). These times are called persistence and exchange times in the condensed matter physics literature [141, 142, 153]. For many dihedrals, we observe that $\langle\tau_{dis}\rangle \ll \langle\tau_{ord}\rangle$ because ($\langle\tau_{dis}\rangle$) is dominated by the short times between transitions in disordered regimes, whereas $\langle\tau_{ord}\rangle$ is heavily influenced by the lengthy times without any transitions in ordered regimes. To calculate these times, CARDS first identifies the time points where a dihedral transitions between two different rotameric states (Fig. 2.2C), referred to as the transition indicator function. The method then extract the complete set of τ_{dis} and τ_{ord} values in the trajectory. Next, CARDS classifies each segment of a trajectory between two consecutive transitions as being in an ordered or disordered regime based on whether the length of the segment t between the transitions is more consistent with the distribution of ordered or disordered times (Fig. 2.2D). We find that transitions within ordered and disordered regimes are roughly Poisson processes with different characteristic times ($\langle\tau_{ord}\rangle$ and$\langle\tau_{dis}\rangle$, respectively), see Fig. A.2. Therefore, CARDS determines if a segment of a trajectory of length t between two consecutive transitions is more

consistent with an ordered or disordered regime using the likelihood ratio (L):

$$L(t) = \frac{P_{dis}(t)}{P_{ord}(t)} = \frac{(\frac{1}{\langle\tau_{dis}\rangle})e^{-\frac{t}{\langle\tau_{dis}\rangle}}}{(\frac{1}{\langle\tau_{ord}\rangle})e^{-\frac{t}{\langle\tau_{ord}\rangle}}} \tag{2.1}$$

where P_{dis} is the probability the segment is disordered and P_{ord} is the probability it is ordered. Taking inspiration from the interpretation of Bayes factors [154], CARDS classifies a segment of a trajectory as being disordered if $L > 3.0$, otherwise the trajectory segment is classified as being in an ordered regime. Our results are robust to varying this cutoff from 1.5 to 5 (Fig. A.1B).

2.3.4 Calculation of structural, disorder-mediated, and holistic correlations

The primary objective of CARDS is to calculate the total correlation between different dihedrals, including both their rotameric state and dynamical state (e.g. whether the dihedral is in an ordered or disordered regime at a given time). Towards this end, we define the holistic correlation (I_H) between dihedrals X and Y as

$$I_H(X,Y) = \overline{I_{ss}(X,Y)} + \overline{I_{dd}(X,Y)} + \overline{I_{ds}(X,Y)} + \overline{I_{sd}(X,Y)} \tag{2.2}$$

where $\overline{I_{ss}(X,Y)}$ is the normalized mutual information between the structure (i.e. rotameric state) of dihedral X and the structure of dihedral Y, $\overline{I_{sd}(X,Y)}$ is the normalized mutual information between the structure of dihedral X and the dynamical state of dihedral Y, $\overline{I_{ds}(X,Y)}$ is the normalized mutual information between the dynamical state of dihedral X and the structure of dihedral Y, and $\overline{I_{dd}(X,Y)}$ is the normalized mutual information between the dynamical state of dihedral X and the dynamical state of dihedral Y. The mutual information (I) is:

$$I(X,Y) = -\sum_{x\in X}\sum_{y\in Y}\frac{P(x,y)}{P(x)P(y)} \tag{2.3}$$

where $x \in X$ refers to the set of possible states that dihedral X can adopt, $p(x)$ is the probability that dihedral X adopts state x, and $p(x, y)$ is the joint probability that dihedral X adopts state x and dihedral Y adopts state y. We define the normalized mutual information ($\overline{I(X,Y)}$) as

$$\overline{I(X,Y)} = \frac{I(X,Y)}{C(X,Y)} \tag{2.4}$$

where $C(X, Y)$ is the maximum possible mutual information between two dihedrals, called the channel capacity [155]. Using this normalized mutual information allows for a direct comparison between the different components of the holistic correlation by correcting for the fact that the largest possible mutual information between different types of dihedrals will vary based on how many different states there are. For example, a side-chain dihedral has three possible rotameric states but only two possible dynamical states, so structural correlations (I_{ss}) can be as large as $log(3)$ while the correlations between dynamical states (I_{dd}) can only reach as high as $log(2)$.

In addition to the above, we define the disorder-mediated correlation (I_{DM}) as all forms of correlation between two dihedrals that rely, at least in part, on disorder ($\overline{I_{sd}} + \overline{I_{ds}} + \overline{I_{dd}}$). This construct is useful for assessing the importance of disorder relative to existing methods based purely on concerted structural changes (I_{ss}). We use bootstrapping to measure the uncertainty in our estimates of all the components of the holistic correlation to ensure any comparisons we make are statistically sound. Specifically, we draw 20 random samples of our trajectories, with replacement, and calculate the structural and disorder-mediated correlations between all pairs of residues. We conclude that disorder-mediated communication dominates if the average disorder-mediated communication minus the standard deviation across all our bootstrap samples is greater than the mean structural correlation plus the standard deviation.

2.3.5 Calculation of net communication to a target site

We are often interested in calculating how much influence a particular residue has over another site, such as an active site or ligand-binding site. To calculate the communication between a reference residue and some target site, we take the average mutual information between two sets of dihedrals: 1) all dihedrals in the reference residue and its nearest neighbors and 2) all dihedrals in the target site. We define the nearest neighbors of a reference residue as all residues with atoms that fall within 3 Å of any atom in the reference residue. Varying this cutoff does not alter our results (Fig. A.1C). Including both a reference residue and its nearest neighbors accounts for the fact that mutating the reference residue will directly change the environment of all neighboring residues.

2.3.6 Calculation of global communication

In addition to identifying residues that have strong correlations to a specific target site, it would also be valuable to identify residues that generally appear to play an important role in allosteric networks. Towards this end, we define the global communication strength of a residue as the sum of its holistic correlations to all other residues. For these calculations, we also include neighboring residues, as in our calculation of the net communication to a target site.

2.4 Results and discussion

2.4.1 Many dihedrals have the potential for disorder-mediated communication

For a dihedral to have ordered and disordered regimes, $\langle \tau_{ord} \rangle$ must be significantly larger than $\langle \tau_{dis} \rangle$. We reasoned that determining if $\langle \tau_{ord} \rangle \geq 3 \times \langle \tau_{dis} \rangle$ is a reasonable heuristic for iden-

tifying dihedrals with separable ordered and disordered regimes based on the likelihood ratio defined in Eq. 2.1. Dihedrals that do not meet this criterion are classified as entirely being in ordered regimes and, therefore, are only capable of having structural correlations with other dihedrals.

Based on the criterion defined above, we find that 556 of the 1584 dihedrals in CAP have separable ordered and disordered regimes and, therefore, are capable of disorder-mediated communication with other dihedrals (Fig. A.3). Mapping these dihedrals to the apo structure of CAP highlights a number of interesting patterns (Fig. 2.3). First of all, CARDS reveals that many side-chain dihedrals buried in CAP's core are capable of disorder-mediated communication. This finding helps to rectify the apparent contradiction between the common physical intuition that proteins' cores should be rigid due to their tight packing and the observation that there is substantial conformational heterogeneity within proteins' cores [57, 156]. That is, dihedrals within a protein's core are commonly locked in a single rotameric state for extended periods of time but rare fluctuations create room for conformational changes. Backbone dihedrals that are capable of disorder-mediated communication tend to reside on the protein's surface. Notably, a number of these backbone dihedrals are in β-sheets that contact cAMP. However, there are also backbone dihedrals within the core that are capable of disorder-mediated communication. For example, we find backbone dihedrals within the central hinge region that are capable of disorder-mediated communication. This observation is noteworthy because the hinge region undergoes a large conformational change upon activation of CAP (Fig. 2.1) [63, 131, 132]. We find that similar patterns emerge when we vary the cutoff for determining whether a dihedral has separable ordered and disordered regimes (Fig. A.4). In the future it will be interesting to examine whether separate proteins, or homologous members of a family, have similar proportions and patterns of dihedrals that are separable into ordered and disordered regimes. However, given the opportunity to only analyze a limited number of sufficiently sampled datasets so far [**?**, 10, 48].

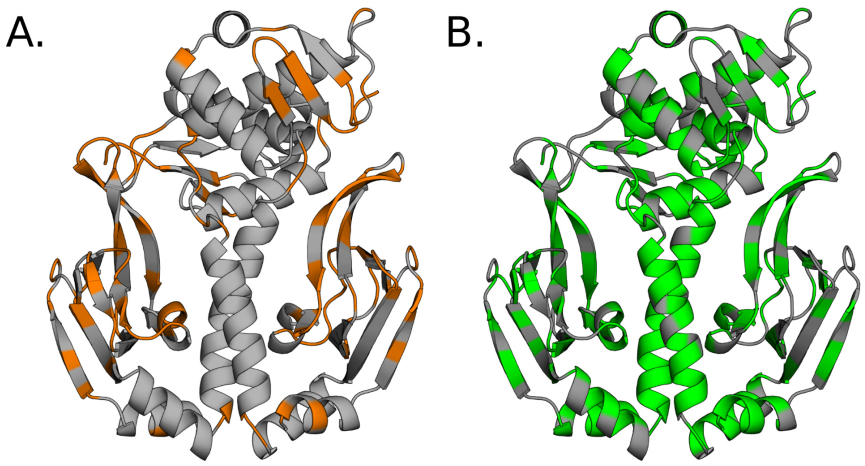


Figure 2.3: Residues whose dihedrals are capable of disorder-mediated communication. **A.** Residues with at least one backbone dihedral that is capable of disorder-mediated communication (orange). **B.** Residues with at least one side-chain dihedral that is capable of disorder-mediated communication (green).

2.4.2 Disorder-mediated correlations dominate communication between the CBDs

Given experimental evidence for allosteric communication between the CBDs without a concerted structural change [63], we expect the disorder-mediated component of the holistic correlation between these sites to be larger than the purely structural component. To test this prediction, we simulated apo CAP for 1.5 μs and calculated the net communication of every residue to one of the cAMP-binding sites. Specifically, we defined the target site as all residues with heavy atoms within 6Åof one of the two cAMP molecules in the holo crystal structure (PDB ID 1CGP) [129]. The residues in this target site are 30, 36, 49, 61-62, 64, 69-86, and 99 from chain A and residues 122-129 from chain B.

As predicted, CARDS successfully identifies that there is communication between the two CBDs. Fig. 2.4A shows the holistic correlations to a single cAMP-binding site. Unsurprisingly, the residues with the strongest correlations to this set of residues reside within the same CBD. However, there is also strong communication between the target site and residues lining the other cAMP-binding pocket. There are also strong correlations on the central hinge region

and the interface between the CBDs and DBDs that may be responsible for allosteric coupling between these domains.

To determine the relative importance of disorder-mediated communication and purely structural correlations, we broke the holistic communication into structural and disorder-mediated components. Furthermore, we used bootstrapping to estimate the uncertainty in each of these components.

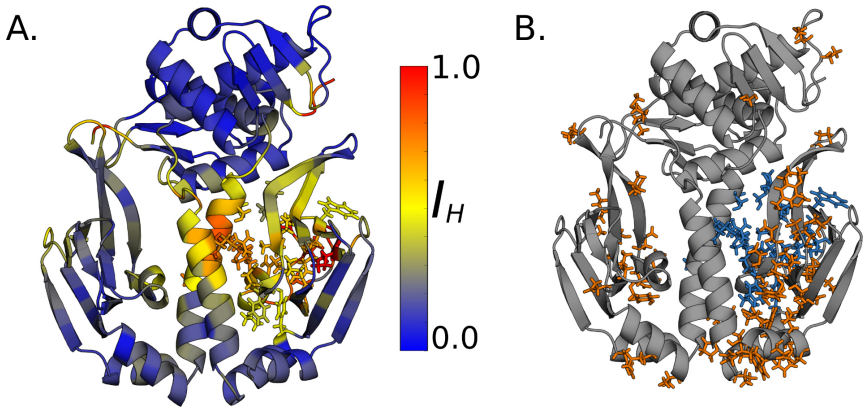


Figure 2.4: Communication to a single CBD. **A.** Holistic mutual information (I_H) of each residue to the residues lining a single cAMP-binding pocket (shown in sticks). **B.** Residues whose communication to the cAMP-binding pocket is dominated by disorder-mediated communication (orange sticks). Residues lining the target cAMP-binding pocket are in blue sticks.

Identifying residues where disorder-mediated communication to the target cAMP-binding site is larger than purely structural correlations automatically identifies a number of residues in the second cAMP-binding site (Fig. 2.4B, Fig. A.5). This finding demonstrates that CARDS recapitulates the experimental finding that communication betwen the two CBDs does not primarily occur through concerted structural changes [63]. CARDS also identifies disorder-dominated communication within a single CBD. Furthermore, the Pearson correlation coefficient between structural and disorder-mediated communication is 0.44. This result indicates there are some similarities between patterns of allosteric coupling that could be observed by focusing entirely on structural correlations and those identified by CARDS, but that considering disorder provides additional information.

2.4.3 Disorder-mediated communication is enhanced in the S62F variant

The S62F variant of CAP is still activated by cAMP binding [63,132,157,158]. However, NMR studies have revealed that the conformation of the DBDs does not change upon cAMP binding. Rather, NMR and ITC experiments suggest an important role for conformational entropy, with the DBDs only changing conformation in the presence of both cAMP and DNA [63, 132]. Therefore, we expect an increase in disorder-mediated communication between the CBDs and DBDs in the S62F variant, compared to wild-type CAP.

To determine the effect of the S62F variant, we also ran 1.5 μs of simulation of this variant. Then we calculated the holistic communication to a single cAMP-binding site, as described for wild-type CAP.

As expected, we observe significant increases in disorder-mediated communication between the target CBD and the DBDs. There are particularly large increases in disorder-mediated correlations in regions of known importance for CAP activation, such as the central hinge region and along the interfaces between the CBDs and DBDs (Fig. 2.5A). At the same time, there are some decreases in disorder-mediated communication within the CBDs. There are also changes in purely structural correlations. These changes often follow the same qualitative trends as the changes in disorder-mediated communication. However, the magnitudes of any increases in purely structural correlations are considerably smaller than the increases in disorder-mediated correlations. Furthermore, reductions in structural correlations are often larger than any decreases in disorder-mediated communication.

2.4.4 Side-chain-side-chain and backbone-side-chain correlations dominate allosteric communication in CAP

To begin understanding the relative importance of different types of degrees of freedom, we plotted the matrix of correlations between every pair of dihedrals (Fig. 2.6A). The upper trian-

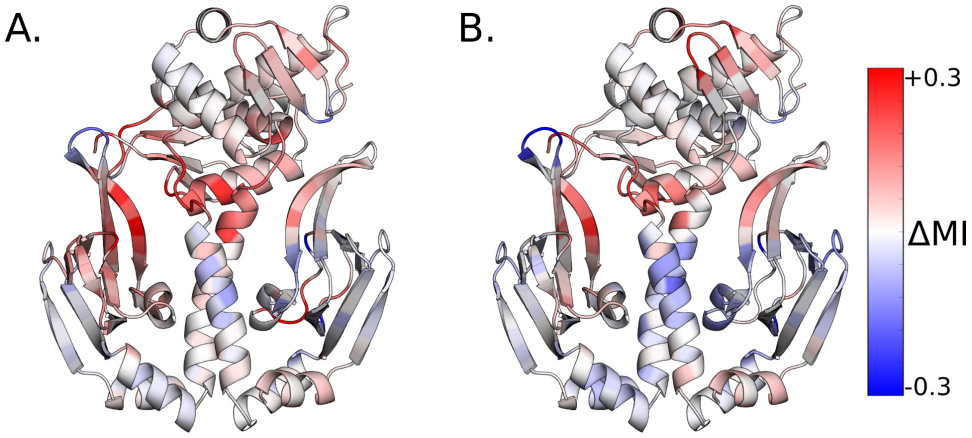


Figure 2.5: Change in coupling to a single CBD pocket upon the S62F mutation. **A.** Change in disorder-mediated communication of each residue to the single cAMP-binding pocket depicted in Fig. 2.4. **B.** Change in structural communication of each residue to the same cAMP-binding pocket. Red indicates increased communication in S62F compared to wild-type, and blue indicates decreased communication.

gle represents purely structural correlations, while the lower triangle represents purely disorder-mediated correlations. Side-chain and backbone dihedrals are also grouped together to enable comparisons between the relative strengths of backbone-backbone, backbone-side-chain and side-chain-side-chain correlations. Inspection of the matrix of all pairwise correlations immediately reveals that side-chain-side-chain correlations dominate allosteric communication in CAP. This observation is consistent with previous reports that side-chain degrees of freedom are more variable than the backbone [57, 156, 159–165]. Backbone-backbone correlations are far rarer, and we find that disorder-mediated correlations between backbone dihedrals are more common than structural correlations. We also find a small number of backbone dihedrals that appear to be hubs of communication that are coupled to the side-chains of a large fraction of the residues in CAP. These hubs appear as lines in the backbone-side-chain quadrants of the matrix in Fig. 2.6A. Mapping the strongest (top 5%) backbone hubs to the structure reveals that they cluster in two regions: the phosphate-binding cassette (PBC) of each CBD and the interface between the CBDs and DBDs, including the central hinge (Fig. 2.6B). Using an even stricter cutoff (top 2%) only identifies residues along the CBD-DBD interface and within the central hinge (Fig. A.6), further emphasizing their importance in the allosteric network. This result suggests that perturbations to these functionally important regions (e.g. cAMP binding)

can influence the behavior of the entire protein, and vice versa.

It is also possible that the assumed Poissonian distribution of dihedral timescales may impact the degree of measured communication. While these are long-tailed distributions, it is not clear if they are truly Poissonian, and proving this behavior may prove a non-trivial task. While we already observe that our measurements are robust to the choice of Likelihood Ratio cutoff (Figure A.1). It will be worth exploring whether or not other exponential distributions can be used to more appropriately model the probability of being ordered or disordered when computing a likelihood ratio.

More importantly, using a different binning strategy than 3-state or 2-state designations can improve measurements of allosteric coupling. These rotameric state designations are based on classic alkane stereochemistry, and assume flat prior. Given the chemical diversity of amino acids, it is possible that some dihedrals will never explore all rotameric states. For example a phenylalanine χ_2 will not explore all three possible rotameric states due to the aromatic ring on the sidechain. One possible avenue to improve the measurement of dihedral communication is to incorporate established rotamer libraries into the rotameric state decomposition [166, 167]. Using library-based designations,and their standard error measurements to define buffer region widths, would correct the currently-used flat prior. This would increase the reliability of measurement of dihedral communication while still allowing us to assess the role of side-chain dihedrals against backbone dihedrals in communication. However, it is important to note that the current binning strategy being used is robust when assessed with metrics such as bootstrapping or Excess Mutual Information. Excess Mutual Information has been shown as a robust way of detecting noise in measuring allosteric communication, and major dihedral communication patterns are still evident after this correction [47]. Consistently, we observe that communication measurements remain significant (ie. not within error of zero) (see Appendix B and Figure A.1). It is worth noting that utilizing this naïve, residue-independent binning system is already capable of making measurable predictions of protein allostery [48, 54].

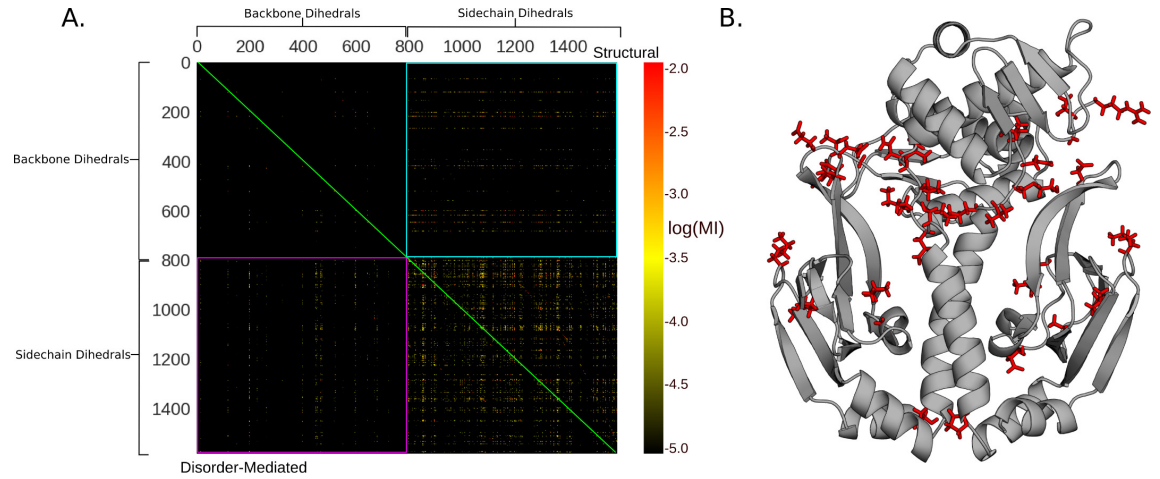


Figure 2.6: Hubs of backbone-side-chain communication in wild-type CAP. **A.** Mutual information between every pair of dihedrals in wild-type CAP. The upper triangle of the matrix (above the green diagonal) represents structural communication and the lower triangle (below the green diagonal) represents disorder-mediated communication. The pink and cyan squares encompass the regions of the matrix that capture backbone-side-chain communication that is mediated by disorder-mediated and structural coupling, respectively. **B.** Structure of wild type CAP highlighting the residues that are the strongest hubs (top 5%) of backbone-side-chain communication (red sticks).

2.4.5 Locating communication hotspots identifies key functional sites

The coincidence of backbone dihedrals that are hubs of communication and key functional sites suggests that CARDS may be capable of predicting the locations of such sites. Indeed, if evolution has selected for communication between particular sites, then one might expect residues in these sites to have stronger coupling to other regions of a protein than typical residues.

To detect strongly communicating residues, we ranked each residue based on the sum of its correlations to all other residues in the protein. This measure of global communication will highlight two types of hotspots: 1) hubs with correlations to many residues and 2) residues that have strong correlations to a few residues.

Coloring each residue in the *apo* structure according to its global communication highlights that the central hinge region and the helices between the two cAMP-binding sites are key mediators of allosteric communication in CAP (Fig. 2.7). There are also hotspot residues in other parts of the cAMP-binding sites and along the interfaces between the CBDs and DBDs.

These are precisely the regions that were identified by assessing communication to a single cAMP-binding site, providing evidence that CARDS can indeed identify key functional sites without foreknowledge of their locations. This conclusion is further supported by the fact that the central hinge region has even stronger communication in the S62F variant (Fig. A.7).

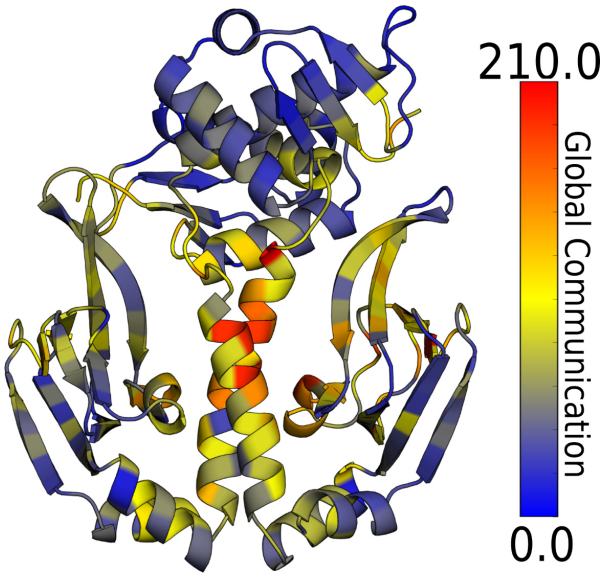


Figure 2.7: Global communication strength of each residue in apo CAP.

2.5 Conclusion

CARDS provides a means to integrate concerted structural changes and disorder-mediated correlations into a holistic view of allostery. Application of this approach to wild-type CAP and the S62F variant demonstrates the method's ability to identify allosteric coupling in the absence of concerted structural changes. Specifically, we showed that examining the coupling of every residue to a known cAMP binding site naturally highlights regions of the protein that are known to be impacted by cAMP binding, such as the second cAMP binding site and the central hinge region connecting the CBDs and DBDs. Decomposing the correlations between these sites into disorder-mediated and purely structural components demonstrates an important role for disorder-mediated coupling in the absence of concerted structural changes. Our global communication metric also provides a means to identify important functional sites without

foreknowledge of their existence and locations. For example, this metric identifies the central hinge region—which undergoes the largest conformational change upon activation—and the cAMP binding pockets as important components of the allosteric network in CAP. Therefore, CARDS should be a powerful means to identify allosteric networks in systems that have not been studied as thoroughly as CAP. Taken together, we expect CARDS to be of great utility for understanding allostery in systems where it is already known to occur, as well as for predicting allostery in systems where it has yet to be observed.

2.6 Acknowledgements

This work was funded by NSF CAREER Award MCB-1552471. G.R.B. holds a Career Award at the Scientific Interface from the Burroughs Wellcome Fund and a Packard Fellowship for Science and Engineering from The David & Lucile Packard Foundation. We are also grateful to NVIDIA Corporation for the GTX Titan X used to run preliminary simulations. Finally, thanks to Kelsey C. Schuster and David Chandler, who, in collaboration with G.R.B., first demonstrated that proteins' side-chains have dynamic heterogeneity.

Chapter 3

Simulation of spontaneous G protein activation reveals a new intermediate driving GDP unbinding

This chapter is adapted from the following publication:

Sun, X. and Singh, S.*, Blumer, K.J., and Bowman, G.R., Simulation of spontaneous G protein activation reveals a new intermediate driving GDP unbinding. eLife, 7, October 2018, https://doi.org/10.7554/eLife.38465.001* [48] **Authors contributed equally to this work*

3.1 Abstract

Activation of heterotrimeric G proteins is a key step in many signaling cascades. However, a complete mechanism for this process, which requires allosteric communication between binding sites that are ∼30 Å apart, remains elusive. We construct an atomically detailed model of G protein activation by combining three powerful computational methods: metadynamics,

Markov state models (MSMs), and CARDS analysis of correlated motions. We uncover a mechanism that is consistent with a wide variety of structural and biochemical data. Surprisingly, the rate-limiting step for GDP release correlates with tilting rather than translation of the GPCR-binding helix 5. β-Strands 1-3 and helix 1 emerge as hubs in the allosteric network that links conformational changes in the GPCR-binding site to disordering of the distal nucleotide-binding site and consequent GDP release. Our approach and insights provide foundations for understanding disease-implicated G protein mutants, illuminating slow events in allosteric networks, and examining unbinding processes with slow off-rates.

3.2 Introduction

Heterotrimeric G proteins are molecular switches that play pivotal roles in signaling processes from vision to olfaction and neurotransmission [168–170]. By default, a G protein adopts an inactive state in which guanosine diphosphate (GDP) binds between the Ras-like and helical domains of the α-subunit (Gα, Fig. 3.1). A dimer consisting of the β- and γ-subunits (G$\beta\gamma$) also binds Gα. G protein-coupled receptors (GPCRs) trigger G protein activation by binding Gα and stimulating GDP release, followed by GTP binding to Gα and dissociation of Gα from G$\beta\gamma$. Gα and $\beta\gamma$ then interact with effectors that trigger downstream cellular responses. Gα returns to the inactive state by hydrolyzing GTP to GDP and rebinding G$\beta\gamma$. Given the central role Gα plays, a common Gα numbering scheme (CGN) has been established to facilitate discussion of the activation mechanisms of different Gα homologs [171]. For example, the notation Lys52$^{G.H1.1}$ indicates that Lys52 is the first residue in helix 1 (H1) of the Ras-like domain (also called the GTPase domain, or G). S6 refers to β-strand 6 and s6h5 refers to the loop between S6 and H5.

Strikingly, the GPCR- and nucleotide-binding sites of Gα are $\sim$30 Å apart (Fig. 3.1), but the allosteric mechanism coupling these sites to evoke GDP release remains incompletely understood [169]. Biochemical and structural studies have elucidated some key steps, but the

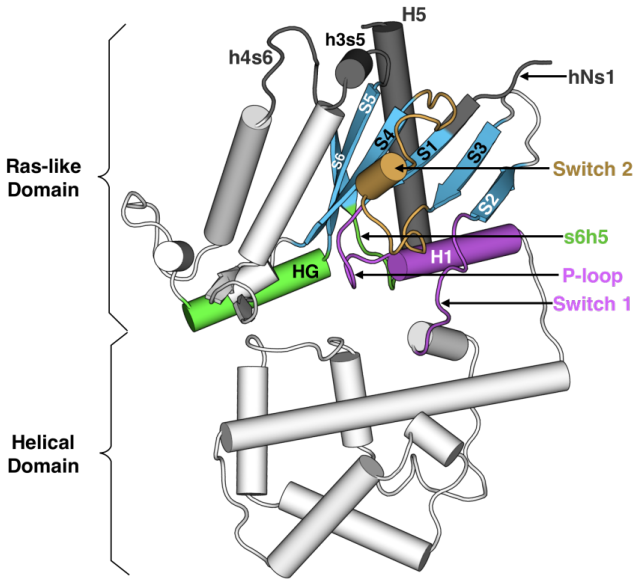


Figure 3.1: Structure of Gαq with key secondary structure elements labeled according to the Common Gα Numbering (CGN) system. The coloring scheme highlights the GPCR binding interface (gray), GDP phosphate-binding regions (pink), GDP nucleotide-binding regions (green), β-sheets (blue), and switch 2 (orange).

entire process has yet to be described in atomic detail. Early studies of Gα subunits revealed structures of the GDP- and GTP-bound states, as well as the transition state for GTP hydrolysis [172–174]. The high similarity of these structures and the binding of GDP or GTP deep in the protein's core suggests that activation occurs by adoption of other conformational states that allow GDP release [175]. One intermediate in G protein activation was suggested by the first crystal structure of a GPCR-bound G protein in which the Ras-like and helical domains of Gα are hinged apart and GDP has dissociated [68]. Structural analysis has led to the proposal of a universal mechanism for G protein activation [171]. In this model, GPCR binding induces translation of H5 away from H1, which increases disorder in H1 and the P-loop (or Walker A motif [173]) to facilitate GDP release. However, there is evidence that additional intermediates may be involved in Gα activation [169, 174, 176, 177],and the functional importance of this conformational ensemble has been previously suggested [178]. Furthermore, mutagenesis and nuclear magnetic resonance (NMR) studies have suggested important roles for other structural elements [179–181].

Molecular dynamics simulations promise to capture the entire mechanism of G protein activation and synthesize the wealth of experimental data on this process. Methodological advances now enable simulations to capture millisecond timescale processes for proteins with less than 100 residues [182]. For example, it is now possible to capture the binding or release of small molecules [47, 99, 183–185] and peptides [186, 187] from small proteins. Impressive simulations on the ANTON supercomputer have revealed critical conformational dynamics of G proteins in their inactive and active states, elucidating the role of domain opening in GDP unbinding [188, 189]. However, even this specialized hardware could not capture the entire process of G protein activation and GDP release due to the size of the Gα subunit (>300 residues) and the slow kinetics of GDP dissociation ($\sim 10^{-3}$ min^{-1}) [190–192].

Here, we introduce an approach to capture rare or long-timescale events, such as GDP release, and reveal the mechanism of Gα activation. As a test of this methodology, we apply it to Gαq, which has one of the slowest GDP release rates [190] and is frequently mutated in uveal melanoma [193, 194]. To highlight aspects of the activation mechanism that we propose are general to all G proteins, we focus our analysis on the behavior of secondary structure elements and amino acids that are conserved across Gα domains. Our approach first combines two powerful sampling methods, metadynamics [195] and Markov state models (MSMs) [196], to observe GDP release and identify the rate-limiting step for this slow process. Then we use our recently developed CARDS method [90], which quantifies correlations between both the structure and disorder of different regions of a protein, to identify the allosteric network connecting the GPCR- and nucleotide-binding sites. The resulting model is consistent with a wealth of experimental data and leads to a number of predictions, described below. Taken together, our results provide the most comprehensive model of G protein activation to date. Based on this success, we expect our approach to be valuable for studying other rare events, including conformational changes and unbinding processes.

3.3 Results and Discussion

3.3.1 Capturing G-protein activation and GDP release in atomic detail

We reasoned that studying the mechanism of spontaneous GDP release from a truncated form of Gαq would be representative of the mechanism of GPCR-mediated activation of the heterotrimeric G protein while minimizing the computational cost of our simulations. This hypothesis was inspired by previous work demonstrating that a protein's spontaneous fluctuations are representative of the conformational changes required for the protein to perform its function [14, 97, 197]. Therefore, we hypothesized that GPCRs stabilize conformational states that G proteins naturally, albeit infrequently, adopt in the absence of a receptor. We chose to focus on Gα since it forms substantial interactions with GPCRs, compared to the relatively negligible interactions between GPCRs and G protein β and γ subunits. This view is supported by the fact that GPCRs and 'mini' G proteins, essentially composed of just the Ras-like domain of Gα, recapitulate many features of GPCR-G protein interactions [198]. We also reasoned that truncating the last five residues of Gαq would facilitate the activation process. These residues contact Gα in some GDP-bound structures but not in GPCR-bound structures [199,200], and removing these residues promotes GDP release due to a reduced GDP-binding affinity [201, 202]. Taken together, such evidence suggests that the last five residues of Gαq help stabilize the inactive state and that removing them would accelerate activation. In support of this hypothesis, we find that the energetic barrier to GDP release is lower in metadynamics simulations of the truncated variant than for full-length Gαq (Figure B.1). These simulations, and those described hereafter, were initiated from an X-ray structure of the Gαq heterotrimer bound to GDP and an inhibitor of nucleotide exchange [203]; G$\beta\gamma$ and the inhibitor were excluded from all simulations.

To observe G-protein activation, we developed a variant of adaptive seeding [204] capable of capturing slow processes like ligand unbinding that are beyond reach of conventional simulation methods. First, we use metadynamics [185, 195, 205] to facilitate broad sampling of

conformational space by biasing simulations to sample conformations with different distances between GDP and Gαq. Doing so forces GDP release to occur but provides limited mechanistic information because adding a biasing force can distort the system's kinetics or cause the simulations to sample high-energy conformations that are not representative of behavior at thermal equilibrium. To overcome these limitations, we chose starting conformations along release pathways observed by metadynamics as starting points for standard molecular dynamics simulations, together yielding an aggregate simulation time of 122.6 μs. These simulations should quickly relax away from high-energy conformations and provide more accurate kinetics. Then we use these simulations to build an MSM (Source data 1). MSMs are adept at extracting both thermodynamic and kinetic information from many standard simulations that, together, cover larger regions of conformational space than any individual simulation [196]. Related approaches have successfully captured the dynamics of small model systems [206] and RNA polymerase [207].

This protocol enabled us to capture the entire mechanism of G-protein activation, including GDP release and the rate-limiting step for this process. Identifying the rate-limiting step for this process is of great value because GDP release is the rate-limiting step for G-protein activation and downstream signaling. Therefore, the key structural and dynamical changes responsible for activation should be apparent from the rate-limiting conformational transition for this dissociation process.

To identify the rate-limiting step, we applied transition path theory [208,209] to find the highest flux paths from bound structures resembling the GDP-bound crystal structure to fully dissociated conformations. Then, we identified the least probable steps along the 10 highest flux release pathways (Figure 3.2A and Figure B.2), which represent the rate-limiting step of release. By comparing the structures before and after the rate-limiting step, we define the bound state as all conformations where the distance from the center of mass of GDP's phosphates to the center of mass of three residues on H1 that contact the GDP β-phosphate (Lys52$^{G.H1.1}$, Ser53$^{G.H1.2}$, and Thr54$^{G.H1.3}$) is less than 8 Å. Consistent with this definition, this distance

remains less than 8 Å throughout the entirety of 35.3 μs of GDP-bound simulations.

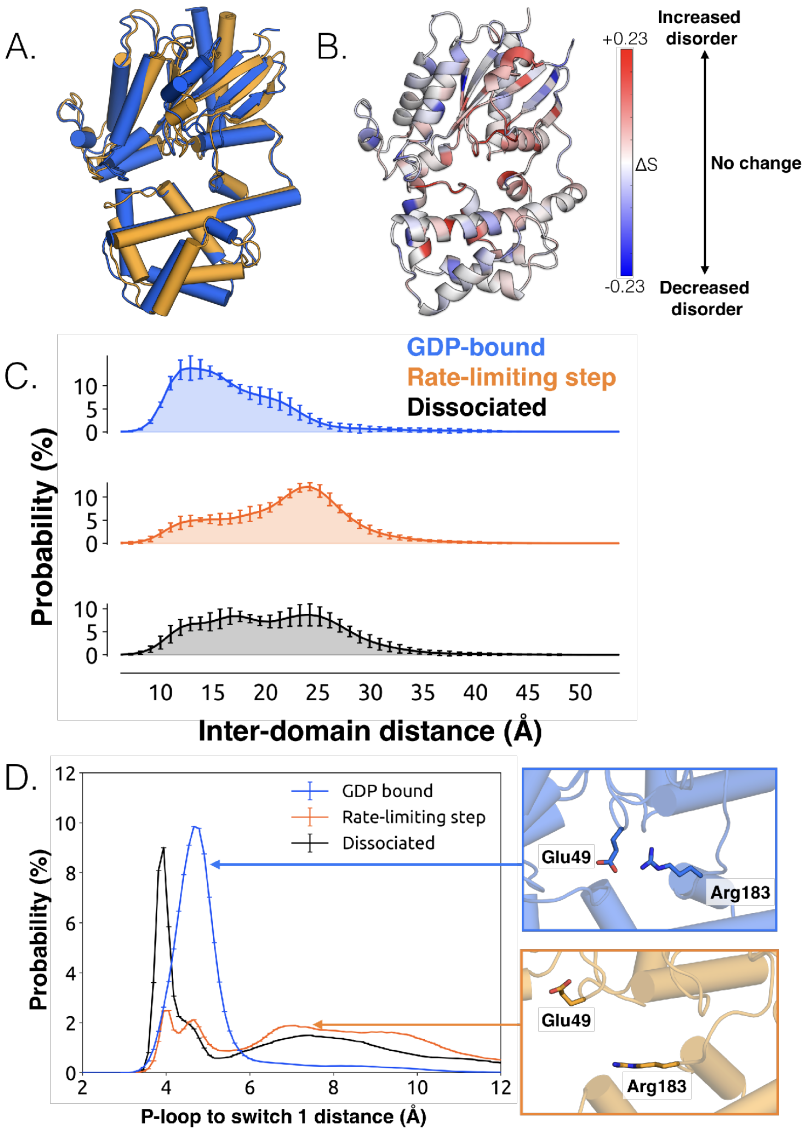


Figure 3.2: Structural and dynamical changes during the rate limiting step for GDP release. **A.** Overlay of representative structures before (blue) and after (orange) the rate limiting step. **B.** Change in conformational disorder (Shannon entropy) across the rate-limiting step, according to the color scale on the right. **C.** Histograms of inter-domain distances before (top, blue) and after (middle, orange) the rate limiting step, along with the inter-domain distance distribution after GDP is released (bottom, black). Inter-domain distance is measured using residues analogous to those used in DEER experiments [188], Leu97$^{H.HA.29}$ in the helical domain and Glu250$^{G.H3.4}$ on H3. **D.** Distribution of distances between Glu49$^{G.s1h1.4}$ and Arg183$^{G.hfs2.2}$ (left) before (blue) and after (orange) the rate-limiting step, as well as after GDP release (black). Representative structures of the interacting residues are also shown (right).

The conformational changes we observe during the rate-limiting step are consistent with pre-

vious structural and biochemical work. For example, we observe that the Ras-like and helical domains separate (Figure 3.2C), as observed in DEER experiments [210] and previous simulations [188]. This finding is consistent with the intuition that these domains must separate to sterically permit GDP release, and that this separation is driven by the disruption of multiple inter-domain interactions. For example, we note a disrupted salt bridge between K275$^{G.s5hg.1}$ and D155$^{H.hdhe.5}$ (Figure B.4), previously identified in structural studies [171]. Domain opening is accompanied by disruption of a key salt bridge between Glu49$^{G.s1h1.4}$ of the P-loop and Arg183$^{G.hfs2.2}$ of switch 1 (Figure 3.2D), as well as an increase in the disorder of many of the surrounding residues (Figure 3.2B and Figure B.3A), consistent with the proposal that this salt bridge stabilizes the closed, GDP-bound state [176].

While domain opening is necessary for GDP release, previous simulations suggest it is insufficient for unbinding [188]. Indeed, we also see that this opening is necessary but not sufficient for GDP unbinding, as the Ras-like and helical domains often separate prior to release (Figure 3.2C). Notably, the Ras-like and helical domains only separate by $\sim$10 Å during the rate-limiting step. In contrast, these domains separate by 56 Å in the first structure of a GPCR-G-protein complex. This result suggests that GDP release may have occurred long before a G protein adopts the sort of widely opened conformations observed in crystallographic structures [68].

3.3.2 Tilting of H5 helps induce GDP release

We also observe less expected conformational changes associated with GDP release. The most striking is tilting of H5 by about 26° (Figure 3.3A, and Figure B.6). We find that H5 tilting correlates strongly with the distance between GDP and Gαq along the highest flux dissociation pathway (Figure 3.3B and B.1). In particular, substantial tilting of H5 is coincident with the rate-limiting step for GDP release. This tilting contrasts with X-ray structures and the universal mechanism, in which H5 is proposed to translate along its helical axis towards the GPCR,

initiating the process of GDP release (Figure 3.3A). During our simulations we also observe translation of H5, but it is not correlated with the rate-limiting step of GDP release (Figure 3.3C, Figure B.7). Therefore, we are not merely missing an important role for translation due to insufficient sampling.

The potential importance of H5 tilting is supported by other structural data. For example, a crystal structure of rhodopsin [211] with a C-terminal fragment from H5 of Gαt shows a similar degree of tilting (Figure 3.3A). Also, the tilt of H5 varies in crystal structures of the β2AR-Gs complex [68], two different GLP-1 receptor-Gs complexes [176, 212], and an A2AR-mini-Gs complex [198]. The potential relevance of tilting has also been acknowledged by a number of recent works [68, 171, 213] including four recently published structures of receptor-G-protein complexes across which H5 also shows a broad range of tilting and translational motion [20–23]. Interestingly, the tilting and translation we observe falls within the observed range of tilting and translational motions that H5 undergoes in available GPCR-G protein complex structures [20–23, 68], providing support that conformational selection plays an important role (B.1). Finally, H5 is translated toward the GPCR in the A2AR-mini-Gs structure but GDP remains bound [198]. The authors of that study originally suggested that one of the mutations in mini-Gs decouples H5 translation from GDP release. However, given that we see GDP release without H5 translation in our simulations, it is also possible that amino acid substitutions required to create mini-Gs instead mitigate H5 tilting. Both of these models are consistent with the fact that some of the mutations in mini-Gs stabilize the GDP-bound state [180].

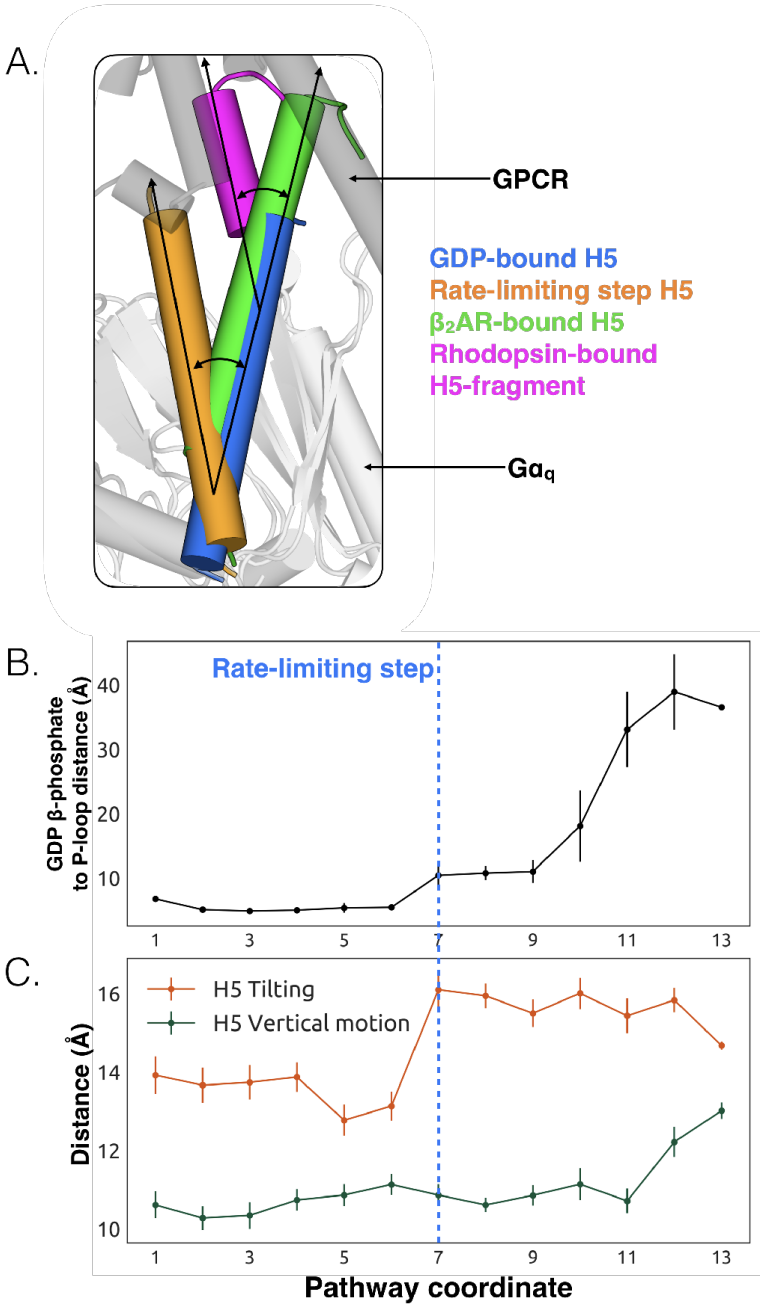


Figure 3.3: Tilting of H5 is correlated with GDP release but translation of H5 is not. **A.** Displacements of H5 relative to the GDP-bound crystal structure (blue). The three other orientations of H5 come from the rate-limiting step in our model (orange), the co-crystal structure of Gαs and β2AR (green, PDB ID 3SN6 [68]), and the co-crystal structure of a C-terminal fragment of Gαt and rhodopsin (magenta, PDB ID 3PQR [211]). The black arrows highlight the change in orientation of the long axis of each helix. A representative GPCR (gray) and Gα (white) structure are shown for reference. **B.** GDP release distance across the highest flux pathway, defined as the distance from the GDP β-phosphate to the center of mass between residues Lys52$^{G.H1.1}$, Ser53$^{G.H1.2}$, and Thr54$^{G.H1.3}$ on H1. The state marking the rate-limiting step is highlighted by the blue dashed line. **C.** H5 motion across the highest flux pathway. The distances measured here representing H5 motion are taken from the same states as in B. H5 tilting (orange) is measured by the distance between Leu349$^{G.H5.16}$ and Tyr325$^{G.S6.2}$. Likewise, H5 vertical motion (green) is measured by the distance between Thr334$^{G.H5.1}$ and Phe341$^{G.H5.8}$. The rate-limiting step is marked with the blue dashed line, extended down from B.

We propose that tilting of H5 is an early step in the GDP release process, which is followed by upward translation of this helix to form a GPCR-G protein complex primed to bind GTP. This hypothesis stems from our observation that tilting of H5 is coincident with the rate-limiting step for GDP release, while translation of H5 only becomes stable after GDP dissociates (Figure 3.3C). This model is consistent with previous suggestions that G-protein activation occurs through a series of sequential interactions with a GPCR [68, 169]. Another possibility is that any displacement of H5, whether tilting or translation, may be sufficient to trigger GDP release.

3.3.3 Identification of the allosteric network that triggers GDP release

While conformational changes of H5 are important for Gα activation, other regions could also play a role [177, 180]. However, it is not straightforward to determine what other structural elements contribute to activation or their importance relative to H5. Our hypothesis that spontaneous motions of a protein encode functionally relevant conformational changes suggests that the coupling between the GPCR- and nucleotide-binding sites of Gα should be present in simulations of the inactive protein; This provides a means to identify key elements of this allosteric network. To test this hypothesis, we ran 35.3 μs of simulation of GDP-Gαq. Then we applied the CARDS method [90] to detect correlations between both the structure and dynamical states of every pair of dihedral angles. Structural states are determined by assigning dihedral angles to the three dominant rotameric states for side-chains (gauche+, gauche-, and trans) and the two dominant rotameric states for backbone dihedrals (cis and trans). Dynamical states are determined by whether a dihedral angle remains in a single rotameric state (ordered) or rapidly transitions between multiple rotameric states (disordered). These pairwise correlations serve as a basis for quantifying the correlation of every residue to a target site, such as the GPCR-binding site. Combining these correlations with structural and dynamical changes in our model of GDP release provides a basis for inferring how perturbations to the GPCR-binding site are transmitted to the nucleotide-binding site. We focus our analysis on the most direct routes for communication between the GPCR- and nucleotide-binding sites by following correlated

motions that emanate from structural elements that directly contact GPCRs until they reach the GDP-binding site. There are correlations between many other elements of Gαq, including parts of the helical domain, that branch off of this allosteric network. Such correlations may be important for aspects of Gα function besides activation, but are beyond the scope of the present study, which focuses on how GPCR-binding impacts nucleotide release.

To understand how H5 tilting triggers GDP release, we first identified structural elements with strong coupling to H5 and then worked our way outward in repeated iterations until we reached the nucleotide-binding site (Figure 3.4). This analysis reveals that tilting of H5 directly communicates with and impacts the conformational preferences of the s6h5 loop, which contacts the nucleobase of GDP (Figure 3.4 and Figure 3.5). During the rate-limiting step, these contacts are broken and there is increased disorder in the s6h5 loop, particularly Ala331 of the TCAT motif (Figure 3.5 and Figure B.3B). The importance of the TCAT motif in our model is consistent with its conservation and the fact that mutating it accelerates GDP release [214–216]. Based on our model, we propose these mutations accelerate release by weakening shape complementarity with GDP.

We also observe an important role for communication from H5 to H1, consistent with the universal mechanism. In particular, H1 is strongly coupled with the s6h5 loop (Figure 3.4B), which is sensitive to displacement of H5. In the rate-limiting step, s6h5 moves away from H1, contributing to an increase in disorder of H1 and the P-loop (Figure B.3A and Figure B.3B). Increased disorder in a set of residues that directly contact the GDP phosphates (Glu49$^{G.s1h1.4}$, Ser50$^{G.s1h1.5}$, Gly51$^{G.s1h1.6}$, Lys52$^{G.H1.1}$, and Ser53$^{G.H1.2}$) likely contributes to a reduced affinity for GDP (Figure B.3A). Increased disorder in these residues also contributes to disruption of the salt bridge between Glu49$^{G.s1h1.4}$ of the Ras-like domain and Arg183$^{G.hfs2.2}$ of the helical domain, facilitating inter-domain separation.

We further note that the s6h5 loop impacts the nucleotide-binding site through allosteric coupling with the HG helix, which also contacts GDP via Lys275$^{G.s5hg.1}$ and Asp277$^{G.HG.2}$ (Fig-

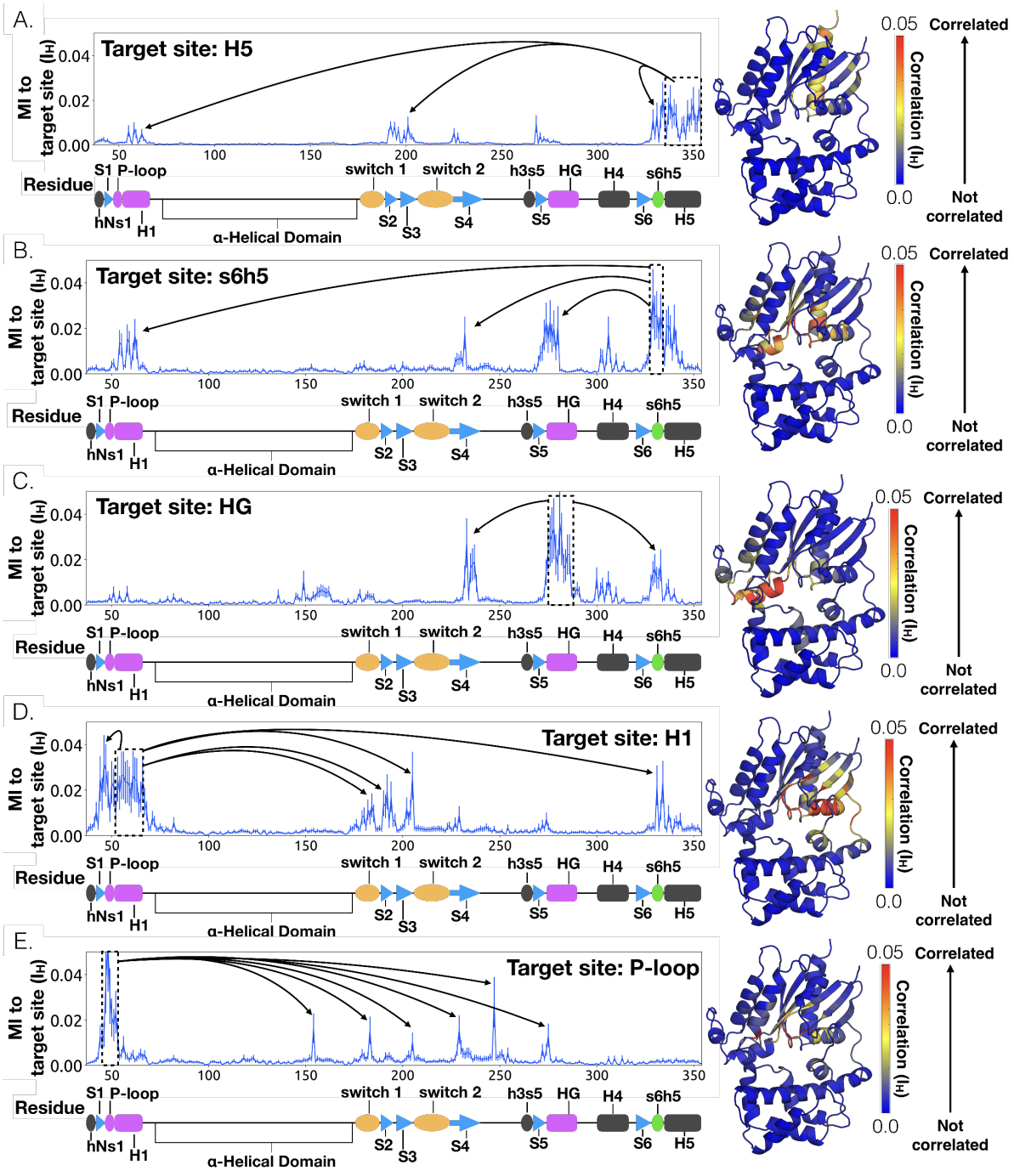


Figure 3.4: Allosteric network connecting H5 motion to the nucleotide binding-site via s6h5. CARDS data showing communication per residue to a target site (dashed box) is plotted (left) and mapped onto the structure of Gαq (right) for **(A)** H5, **(B)** s6h5, **(C)** HG, **(D)** H1, and **(E)** the P-loop. Arrows indicate regions of importance with significant communication to the target site.

ures 3.4E and 3.6). The disorder of both of these residues increases during the rate-limiting step, consistent with observations of increased mobility in HG upon receptor-mediated ac-

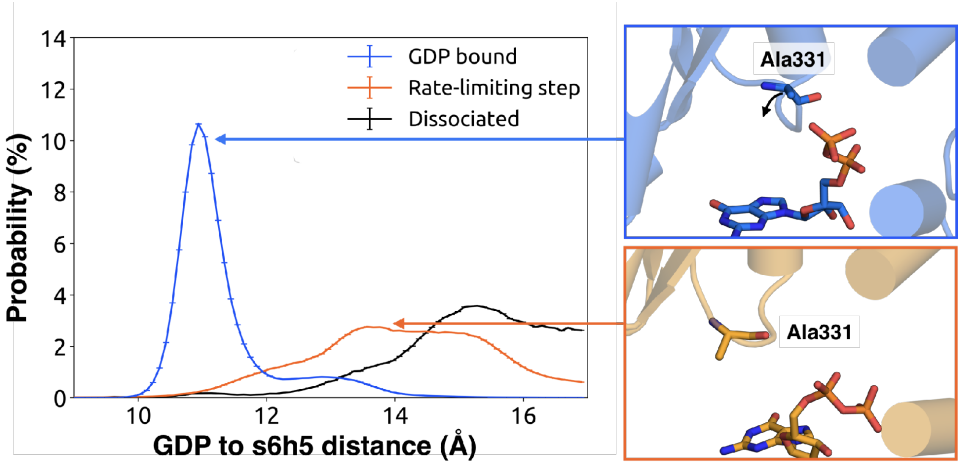


Figure 3.5: Change in the s6h5 loop conformation across the rate-limiting step. Distribution of distances (left) from GDP's β-phosphate to Ala331$^{G.s6h5.3}$ on the s6h5 loop before (blue) and after (orange) the rate-limiting step, as well as after GDP release (black). Representative structures of the s6h5 loop (right) are shown for before (top right, blue) and after (bottom right, orange) the rate limiting step.

tivation (Oldham and Hamm, 2008). There are also correlations between the P-loop and Lys275$^{G.s5hg.1}$ on Helix G (Figure 3.4E), which result from the disruption of a key salt bridge between Lys275$^{G.s5hg.1}$ and Glu49$^{G.s1h1.4}$ on the P-loop during the rate-limiting step (Figure 3.6). Lys275$^{G.s5hg.1}$ is conserved across all Gα families, suggesting it plays an important role in the stability or function of the protein. However, attempts to experimentally examine the role of this residue by mutating Lys275$^{G.s5hg.1}$ have resulted in aggregation [180]. Our simulations suggest Lys275$^{G.s5hg.1}$ plays an important role in stabilizing the GDP-bound state and that breaking the salt bridge with Glu49$^{G.s1h1.4}$ facilitates GDP release. This finding demonstrates the utility of our atomistic simulations, as we can examine the role of residues that are difficult to probe experimentally.

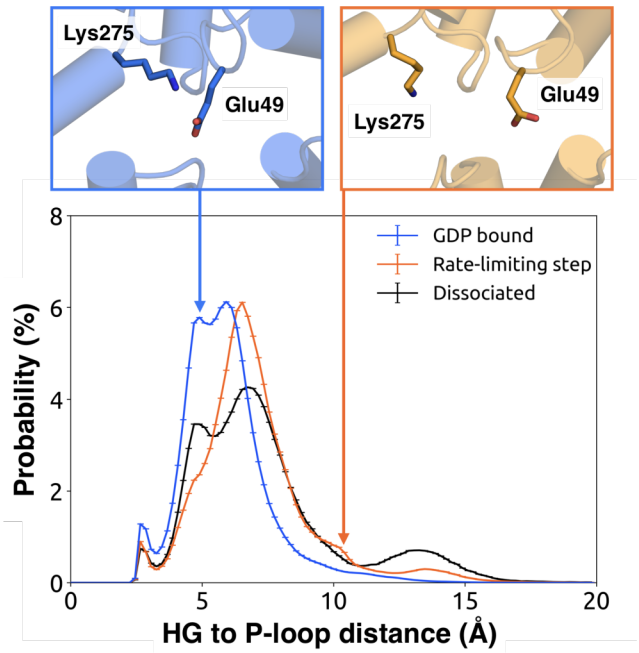


Figure 3.6: Probability distributions of the distance between the side-chains of Lys275$^{G.s5hg.1}$ and Glu49$^{G.s1h1.4}$. Distributions were computed for the bound (blue), rate-limiting step (orange), and dissociated (black) states. Representative structures (above) for the bound (left, blue) and rate-limiting step (right, orange) are included with residues as sticks.

3.3.4 H1 and β-sheets are communication hubs

To identify other important structural elements in the allosteric network underlying G protein activation, we followed correlated motions emanating from other sites that are known to interact directly with GPCRs, including the hNs1 loop, the h3s5 loop, and the h4s6 loop [169]. We find that h3s5 and h4s6 are largely isolated, suggesting they play a role in forming a stable GPCR-G protein complex but not in the allosteric mechanism that triggers GDP release. This finding is consistent with sequence analysis suggesting these structural elements are important determinants of the specificity of GPCR-Gα interactions [217]. In contrast, the hNs1 loop has strong correlations with β-strands S1-S3 (Figure 3.7). These strands, in turn, communicate with H1, switch 1, and the P-loop.

Integrating our correlation analysis with structural insight from the rate-limiting step described above suggests an important role for S1-S3 in a complex allosteric network that couples the

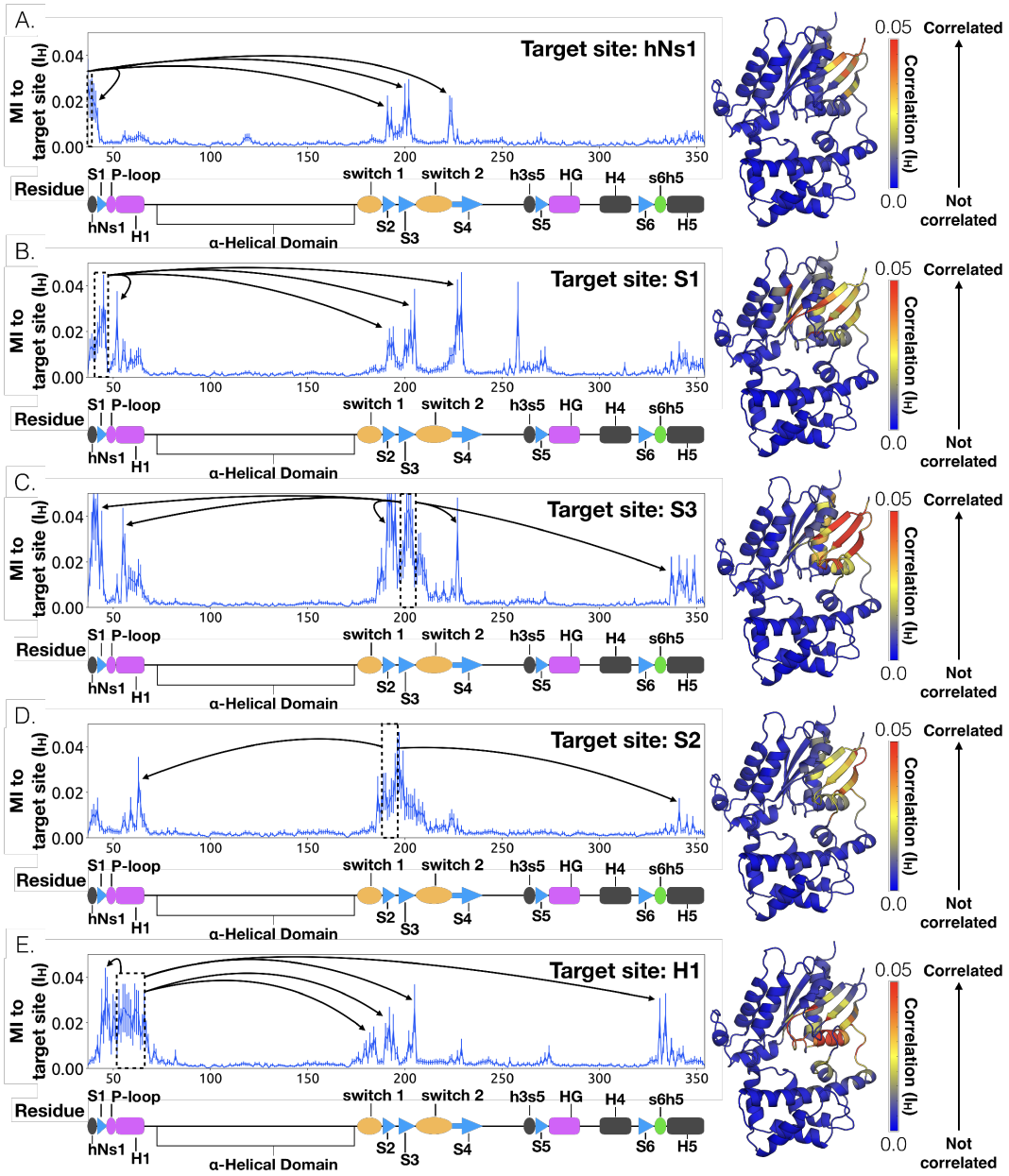


Figure 3.7: CARDS data showing communication per residue to a target site (dashed box) is plotted (left) and mapped onto the structure of Gαq (right) for **(A)** hNs1, **(B)** S1, **(C)** S3, **(D)** S2, and **(E)** H1. Arrows indicate regions of importance with significant communication to the target site.

GPCR- and nucleotide-binding sites (Figures 3.7 and 3.8, Figure B.8). S2 and S3 twist relative to S1 and away from H1 (Figures 3.2A and 3.9, Figure B.10). This twisting disrupts stacking between Phe194$^{G.S2.6}$ on S2 and His63$^{G.H1.12}$ on H1 and increases disorder of side-chains in H1 (Figures 3.2B and 3.9, Figure B.3C). Increased disorder in H1 is also a crucial component of the proposed universal mechanism, but in that model translation of H5 is the key trigger

for changes in H1. The role for the β-sheets in our model is consistent with previous work identifying interactions between S2 and H1 [171], NMR experiments showing chemical exchange in the methyls of S1-S3 upon receptor binding [179], and mutational data. In particular, Flock et al. have previously noted the important interaction between residues Phe194$^{G.S2.6}$ and His63$^{G.H1.12}$ [171].

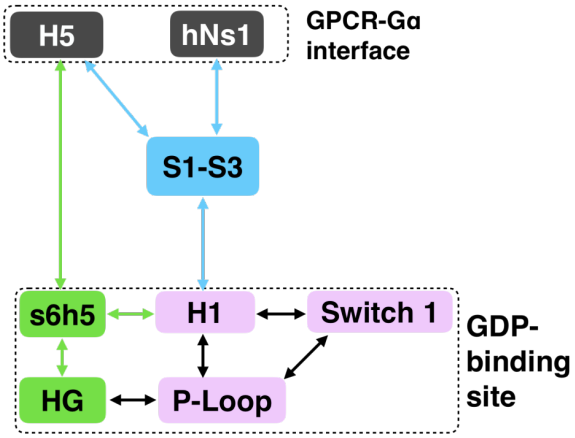


Figure 3.8: Allosteric network connecting the GPCR- and nucleotide-binding interfaces. The coloring scheme corresponds to that used in 3.1, highlighting the GPCR binding interface (gray), GDP phosphate-binding regions (pink), GDP nucleotide-binding regions (green), and the β-sheets (blue).

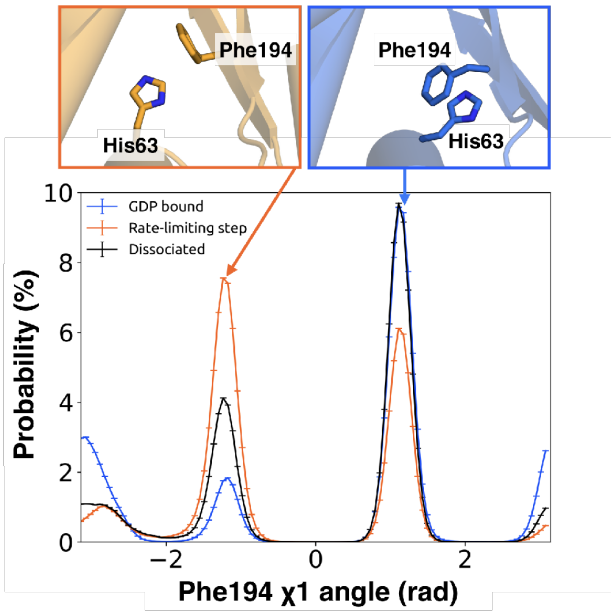


Figure 3.9: π–stacking between S2 and H1 is disrupted during the rate-limiting step. Distribution of the χ_1 angle (bottom) of Phe194$^{G.S2.6}$ on S2 before (blue) and after (orange) the rate-limiting step, as well as after GDP release (black). Representative structures of Phe194$^{G.S2.6}$ and His63$^{G.H1.12}$ (top) corresponding to before and after the rate-limiting step are also shown.

The importance of H1 and β-strands S1-S3 is underscored by mapping the global communication of every residue onto a structure of Gα (Figure B.9). The global communication of a residue is the sum of its correlations to every other residue and is a useful metric for identifying residues that are important players in allosteric networks [90]. Interestingly, these β-strands and H1 have higher global communication than H5 and the s6h5 loop. This suggests that H1 and the β-sheets integrate conformational information from multiple sources, including the hNs1 loop, and not just H5. The importance of the β-sheets and H1 for allosteric communication is consistent with their conservation [180], which may not simply reflect the role they play in protein folding and stability, as had been suggested previously [180, 218].

3.3.5 GDP release alters the structure and dynamics of the Gβ-binding site

We also find that switch 2 has strong correlations with the nucleotide-binding site, especially switch 1 (Figure B.8). Given that switch 2 is a major component of the interface between Gα and Gβ, this communication could enable GDP release to trigger dissociation of Gα from G$\beta\gamma$. Examining the rate-limiting step for GDP release reveals that switch 2 shifts towards the nucleotide-binding site (Figure 3.10) and exhibits increased conformational disorder (Figure 3.2B and Figure B.3). These findings are consistent with previous kinetic studies postulating that switch 2 dynamics are impacted prior to GDP release [219].

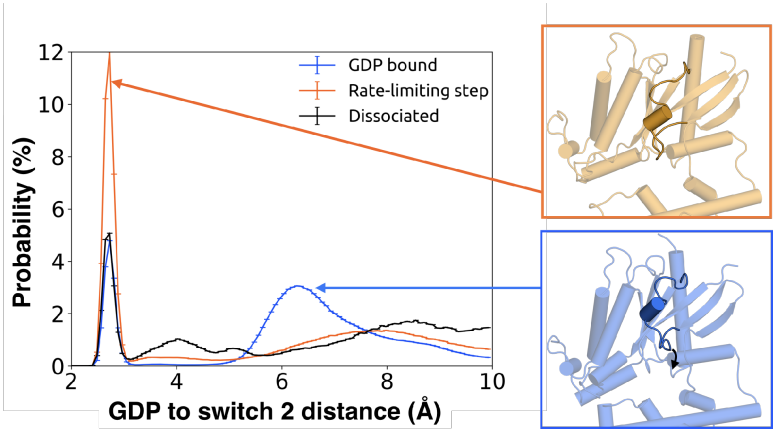


Figure 3.10: Switch 2 moves towards GDP across the rate-limiting step. Distance distribution (left) of Gly207$^{G.s3h2.1}$ to GDP before (blue) and after (orange) the rate-limiting step, as well as after GDP release (black). Representative structures of Switch 2 motion are shown (right).

3.4 Conclusion

We have succeeded in simulating G protein activation, including both the allosteric coupling between the GPCR- and nucleotide-binding sites of Gαq and consequent unbinding of GDP. Our results reveal a previously unobserved intermediate that defines the rate-limiting step for GDP release and, ultimately, G protein activation. Our model synthesizes a wealth of experimental data and previous analyses. For example, we identify an important role for coupling

from H5 to the s6h5 loop and H1 that is consistent with a previously proposed universal mechanism for G protein activation. However, we also find that this allosteric network incorporates the hNs1 loop, β-strands S1-S3, and the HG helix. Strands S1-S3 and H1 serve as hubs in this network, simultaneously integrating information from both H5 and the hNs1 loop. Our observation is consistent with previous postulates that information flows from H5 and hNs1 to H1 [220]. It is important to note that our model was extracted using simulations of Gαq, and so some correlations or changes in conformation and dynamics may apply only to Gαq. However, by focusing our analysis on secondary structure elements and residues that are shared across all Gα homologs, our model likely captures a universal 'skeleton' of changes involved in Gα activation, expanding upon a previously proposed universal mechanism for Gα activation [171]. The consistency of our model with a wide variety of structural and biochemical data suggests that it is a promising foundation for future efforts to understand the determinants of GPCR-Gα interaction specificity, how mutations cause aberrant signaling and disease, and how small molecule inhibitors modulate Gα activation. Our model also adds weight to the growing appreciation for the fact that a protein's spontaneous fluctuations encode considerable information about its functional dynamics [20–23]. Construction of our model was enabled by a powerful combination of simulation methods, namely metadynamics and MSMs. In the future, we expect this methodology will prove valuable for understanding other slow conformational changes and unbinding processes.

3.5 Materials and Methods

3.5.1 Molecular dynamics simulations of GDP unbinding

System preparation

We used the crystal structure of Gα from the heterotrimeric Gαq protein (PDB entry 3AH8) as the initial structure to set up our simulations [203]. The first 36 residues were removed prior to simulation, as they come from Gαi, along with the Gβ and Gγ subunits which were removed to minimize the system size and maximize our chances of observing GDP release [169].

The protein structure was solvated in a dodecahedron box of TIP3P water [143] that extended 1 nm beyond the protein in every dimension. A single Mg2+ ion was added coordinating the phosphate groups of GDP as its presence accelerates GDP release [221]. Thereafter, Na+ and Cl− were added to produce a neutral system at 0.15 M NaCl. The final system consists of one GDP, one Gαq, 57 Cl^{-1}, 64 Na^{+1}, one Mg^{2+}, and 18696 TIP3P water molecules, for a total of 61,893 atoms.

Molecular dynamics (MD) simulations were carried out using Gromacs [145, 222] and the AMBER03 force field [146]. The force field parameters of GDP were obtained from the AMBER Parameter Database (http://research.bmh.manchester.ac.uk/bryce/amber) [223]. The system was energy minimized with the steepest descent algorithm for 50,000 steps until the maximum force fell below 100 kJ/mol/nm using a step size of 0.02 nm and a cut-off distance of 0.9 nm for the neighbor list, electrostatic interactions and van der Waals interactions. Afterward, the solvent was relaxed by a NVT simulation of 100 ps with the constraint of 1000 kJ/mol/nm applied to the protein heavy atoms and 2 fs as the time step. In this relaxation simulation, the temperature of the system is constrained to 300 K using V-rescale thermostat (with a time constant of 0.1 ps) [224]. A cut-off distance of 1 nm was used for the van der Waals, short-range electrostatic interactions. Periodic boundary conditions are applied to x, y and z directions. The

Particle-Mesh-Ewald method is employed to recover the long-range electrostatic interactions with 0.16 nm as the grid spacing and with a fourth order spline [225]. All the bonds connecting to hydrogens are constrained using LINCS algorithm [226]. After the NVT relaxation, the system further underwent an NPT simulation for one ns with the time step of 2 fs for equilibration. The simulation parameters here are kept the same as the NVT relaxation except for the application of Parrinello-Rahman barostat for pressure coupling [227]. After these equilibration runs, the constraint on heavy atoms were removed for all subsequent production runs. Virtual sites were used to allow for a 4 fs time-step [228]. We then applied a three-step protocol to simulate GDP release.

Step one: MD simulations of the GDP-bound state

We performed 100 parallel simulations of the GDP bound state on the Folding@home [37] distributed computing environment for an aggregate simulation time of 35.3 μs.

Step two: Metadynamics simulations

We subsequently ran metadynamics simulations [195, 229] initiated from conformations generated in step one to actively promote GDP release. Starting conformations were selected from step one by clustering protein conformations into 625 states using a hybrid K-center/K-medoids method [230] with a 2 Åcutoff. The 10 most populated states included conformations with large inter-domain separation distances between the Ras-like and AH domains as measured by the angle formed between the α-carbon atoms of Leu97$^{H.HA.29}$, Asn82$^{H.HA.14}$, Ile258$^{G.H3.12}$, and Glu250$^{G.H3.4}$. This inter-domain angle ranged from $-30°$ to $0°$. From these states, three representative structures were chosen with inter-domain angles of $-6°$ (Conf. 1),$-10°$ (Conf. 2), and $-26.6°$ (Conf. 3) as starting conformations for metadynamics simulations, which were run on PLUMED [229]. We defined two collective variables for our metadynamics simulations: 1) the distance between GDP's phosphate groups and the backbone of Lys52$^{G.H1.1}$-Thr54$^{G.H1.3}$ in

Table 3.1: Details of metadynamics simulations

Starting conformation	Width of CV1 (Å)	Width of CV2 (Å)	Number of conformations selected
1	0.1	0.1	171
1	0.08	0.03	132
1	0.03	0.01	504
2	0.1	0.1	145
2	0.08	0.03	141
2	0.03	0.01	320
3	0.1	0.1	198
3	0.08	0.03	186
3	0.03	0.01	288

Gαq subunit (CV1), and 2) the RMSD of GDP to the starting conformation (CV2).

In metadynamics, a history-dependent biased potential VG(S,t) is added to the two selected CVs.

$$V_G(S,t) = \int_0^t (dt'\omega \exp(-\sum_{i=1}^{d} \frac{(S_i(R) - S_i(R(t')))^2}{2\sigma_i^2})) \tag{3.1}$$

where t represents time, S are collective variables, ω is the energy rate and σ_i controls the width of the Gaussian for the ith collective variable. Summing up the Gaussians allows us to obtain the biased potential V_G. The free energy $-F(S)$ is derived by the assumption,

$$\lim_{t\to\infty} (S,t) \sim -F(S) \tag{3.2}$$

The metadynamics simulations were repeated three times for each selected representative structure using different Gaussian widths (3.1). We set the Gaussian height to 1.5 kJ/mol.

We observed the release of GDP from Gα in the metadynamics simulations and obtained the free-energy landscape for the release. We then applied the string method [231] to detect potential release pathways and use these conformations as the seeds for simulations in step 3. We can use the function $\chi(\alpha)$ to represent $\chi(C1, C2)$ which showing the minimum free-energy path.

We thus have

$$\frac{dZ(\alpha)}{d\alpha} = \sum_{k=1}^{n} \frac{\partial \theta_i(\chi(\alpha))}{\partial \chi_k} \frac{d\chi_k}{d\alpha} \tag{3.3}$$

which is parallel to

$$\sum_{k=1}^{n} \frac{\partial \theta(\chi(\alpha))}{\partial \chi_k} \frac{dF(\chi_k)}{d\chi_k} = \sum_{j,k=1}^{n} \frac{\partial \theta_i(\chi(\alpha))}{\partial \chi_k} \frac{\partial \theta_j(\chi(\alpha))}{\partial \chi_k} \frac{dF(z(\alpha))}{dZ(\alpha)} \tag{3.4}$$

The average of the tensor

$$\sum_{j,k=1}^{n} \frac{\partial \theta_i(\chi(\alpha))}{\partial \chi_k} \frac{\partial \theta_j(\chi(\alpha))}{\partial \chi_k} \tag{3.5}$$

can be represented as

$$M_{ij}(z) = \Omega^{-1} e^{\beta F(z)} \int_{\mathbb{R}^N} \sum_{k=1}^{n} \frac{\partial \theta_i(\chi(\alpha))}{\partial \chi_k} \frac{\partial \theta_j(\chi(\alpha))}{\partial \chi_k} e^{-\beta F(z)} \tag{3.6}$$

$$\prod_{v=1}^{N} \delta(z_v - \theta_v(\chi)) d\chi \tag{3.7}$$

The points determining the minimum free-energy path along the surface satisfy

$$0 = (M_{ij}(z) \Delta F(z(\alpha)))^{\perp} \tag{3.8}$$

We applied this method [232, 233] to obtain a minimum free energy path, extracting 2085 conformations along potential GDP release pathways. Notably, we only observed the transition from the bound state and the unbound state for one time in a single metadynamics simulation. This implies that the predicted free-energy surface from metadynamics cannot be used to describe the release events accurately. In spite of this, we can still use the conformations along the release pathway explored by metadynamics to seed unbiased parallel MD simulations.

Step three: Metadynamics-seeded MD simulation of GDP release

Lastly, we carried out unbiased MD simulations initiated from the 2085 conformations obtained from metadynamics using the Folding@home platform [37]. A total of 122.6 μs of simulation was generated in this step. All the following analyses are based on unbiased MD simulations.

3.5.2 Identifying the allosteric network with CARDS

To determine how the GPCR- and GDP-binding regions communicate with one another, we applied the CARDS [90] methodology to simulations of the GDP-bound state of Gαq. CARDS measures communication between every pair of dihedrals via both correlated changes in structural motions and dynamical behavior. Structural states are captured by discretizing backbone ϕ and ψ dihedrals into two structural states (cis and trans), while side-chain χ angles are placed into three states (gauche+, gauche-, and trans). Every dihedral is also parsed into dynamical states, capturing whether the dihedral is stable in a single state (ordered), or rapidly transitioning between multiple states (disordered). These dynamical states are identified using two kinetic signatures of dihedral motion: the average time a dihedral persists in a structural state (an ordered timescale), and the typical timescale for transitions between structural states (a disordered timescale). Parsing into dynamical states utilizes a two-step process by (i) calculating the distribution or ordered and disordered times from the simulations and (ii) assigning each period of time between two consecutive transitions into ordered and disordered states based on which distribution the time between two transitions is most consistent with.

From these states, a holistic communication ($I_H(X,Y)$) is computed for every pair of dihedrals X and Y:

$$I_H(X,Y) = \overline{I_{ss}(X,Y)} + \overline{I_{dd}(X,Y)} + \overline{I_{ds}(X,Y)} + \overline{I_{sd}(X,Y)} \tag{3.9}$$

where $\overline{I_{ss}(X,Y)}$is the normalized mutual information between the structure (i.e., rotameric state) of dihedral X and the structure of dihedral Y, $\overline{I_{ds}(X,Y)}$ is the normalized mutual infor-

mation between the structure of dihedral X and the dynamical state of dihedral Y, $\overline{I_s d}(X, Y)$ is the normalized mutual information between the dynamical state of dihedral X and the structure of dihedral Y, and $\overline{I_{dd}(X, Y)}$ is the normalized mutual information between the dynamical state of dihedral X and the dynamical state of dihedral Y. The Mutual Information (I) is

$$I(X, Y) = -\sum_{x \in X} \sum_{y \in Y} \frac{P(x, y)}{P(x)P(y)} \tag{3.10}$$

where $x \in X$ refers to the set of possible states that dihedral X can adopt, $p(x)$ is the probability that dihedral X adopts state x, and $p(x, y)$ is the joint probability that dihedral X adopts state x and dihedral Y adopts state y. Normalized mutual information is computed using the maximum possible mutual information for any specific mode of communication.

From the pairwise correlation for every dihedral-pair, we extracted how much each individual residue communicates with a target site of interest via bootstrapping with 10 random samples with replacement. After locating the group of residues communicating most strongly with a specific target site, we set this newly identified group as the new target site; The iteration of this process allows us to identify a pathway of communication from one region of interest to another. Here, we set the GPCR contact sites as our initial target sites. We then iteratively used this approach to identify pathways connecting these contact sites with the GDP-binding site of Gαq.

3.5.3 Markov state model construction

We clustered Gαq conformations and GDP binding states separately and combined the assignments to build a Markov State Model using MSMbuilder [230, 234] and enspara [39]. First, we clustered protein conformations into 5040 states using a hybrid k-center/k-medoids method with 1.8 Å cutoff. Then we clustered the GDP-binding state into 321 states using the automatic partitioning algorithm (APM) [235] with a residence time of 2 ns. By combining the assign-

ments from protein conformations and the GDP-binding states, we obtained a total of 221,965 states. The implied timescales of this MSM show Markovian behavior with a lag time of 5 ns [236]. (Figure B.5). Analyses of distances, angles, and dihedrals of interest were carried out using bootstrapping with ten random samples, with replacement. Results were insensitive to varying the number of bootstrapped samples between 5 and 30. Histograms were generated using 100 bins.

3.5.4 Quantifying conformational disorder

The disorder of every residue was measured by computing Shannon entropy [237] of each dihedral as they are natural degrees of freedom for describing protein dynamics. Shannon entropy (H) is defined as

$$H(X) = -\sum_{x \in X} p(x) log(p(x)) \tag{3.11}$$

where $x \in X$ refers to the set of possible states that dihedral X can adopt and $p(x)$ is the probability that dihedral X adopts state x. Dihedral angles were calculated using MDTraj (McGibbon et al., 2015) and assigned to discrete rotameric states as described above using CARDS. The entropy of a single residue was computed by summing up the entropies of its dihedrals, and normalized by the residue's maximum possible Shannon entropy. This maximum possible Shannon entropy, using a flat distribution for the appropriate number of bins, is referred to as the 'channel capacity' and has been used to normalize other information-theoretic metrics [90]. Summing entropies within a residue establishes an upper bound on the degree of motion for a single residue, while ignoring intra-residue correlations between dihedrals.

3.5.5 Identification of the rate-limiting step for GDP release

We used transition path theory (TPT) [208, 209] to find the highest flux paths from the bound state to the unbound state [238]. The bound state was defined as all clusters that satisfied two

criteria: (i) GDP is within 6 Å of the backbone atoms of Lys52$^{G.H1.1}$-Thr54$^{G.H1.1}$, and (ii) GDP has an RMSD $< 0.5\mathring{A}$ to its crystallographic conformation. The unbound state was defined as all clusters with GDP $> 55\mathring{A}$ from the binding pocket. The rate-limiting step was identified by finding the bottleneck in the highest flux paths. To obtain this, we first calculate the flux between states i and j along any possible unbinding path using

$$f(x) = \begin{cases} \pi_i q_i^- T_{ij} q_j^+, & \text{if } i \neq j \\ 0, & \text{if } i = j \end{cases}$$

where q^+ is the committor probability from the bound to the unbound state, and q_i^- is $1 - q_i^+$; π_i is the weighted probability, and T_{ij} is the transition matrix. The highest flux paths can be identified by maximizing the fluxes between the bound states and the unbound states using

$$c(w) = \min\left(f^+_{i_l i_{l+1}} | l = 1 \ldots n_w - 1\right) \tag{3.12}$$

where i_l are intermediate states. From this, the slowest step was extracted as the minimum flux step of the highest flux release pathway.

3.6 Acknowledgements

We thank TE Frederick and TD Todd for their helpful discussion and insight. We are grateful to the Folding@home users for computing resources.

Chapter 4

Discovery of a cryptic allosteric site in Ebola's 'undruggable' VP35 protein using simulations and experiments

This chapter is adapted from the following publication:

Cruz, M.A. and Frederick, T.E.*, Singh, S., Vithani, N., Zimmerman, M.I., Porter, J.R., Moeder, K.E., Amarasinghe, G.K., and Bowman, G.R., Discovery of a cryptic allosteric site in Ebola's 'undruggable' VP35 protein using simulations and experiments. Preprint on BioRxiv https://doi.org/10.1101/2*
[54]

**Authors contributed equally to this work*

In this work, my work analyzing the allostery in VP35 and the coupling between the cryptic pocket opening and binding interface is presented in figure 4.3 and appendix figure C.2.

4.1 Abstract

Many proteins are classified as 'undruggable,' especially those that engage in protein-protein and protein-nucleic acid interactions. Discovering 'cryptic' pockets that are absent in available structures but open due to protein dynamics could provide new druggable sites. Here, we integrate simulations and experiments to search for cryptic pockets in Ebola viral protein 35 (VP35). VP35 plays essential roles in Ebola's replication cycle, including binding the viral RNA genome to block a host's innate immunity. However, VP35 has so far proved undruggable. Using adaptive sampling simulations and allosteric network detection algorithms, we uncover a cryptic pocket that is allosterically coupled to VP35's key RNA-binding interface. Experimental tests corroborate the predicted pocket and confirm that stabilizing the open form allosterically disrupts RNA binding. These results demonstrate simulations' power to characterize hidden conformations and dynamics, uncovering cryptic pockets and allostery that present new therapeutic opportunities.

4.2 Introduction

Many proteins have proved so difficult to target with small molecule drugs that they are often classified as undruggable, greatly limiting the scope of drug design efforts. In fact, up to 85% of human proteins have been classified as undruggable because their folds are thought to lack binding pockets where small molecules can bind with the affinity and specificity required for drug design [239]. Many undruggable proteins predominantly participate in protein-protein interactions (PPIs) and protein-nucleic acid interactions (PNIs) [72, 240]. In contrast to the binding pockets that many enzymes and receptors use to bind their small molecule ligands, the large flat interfaces involved in PPIs and PNIs do not lend themselves to forming many favorable interactions with small drug-like molecules. As a result, PPIs and PNIs are often considered intractable targets even when there is strong evidence that disrupting these interac-

tions would be of great therapeutic value.

Cryptic pockets could provide new opportunities to target undruggable proteins [41, 241], but realizing this potential remains challenging. Such pockets are absent in available experimental structures because they only form in a subset of excited states that arise due to protein dynamics. Cryptic sites can serve as valuable drug targets if they coincide with key functional sites, as can cryptic allosteric sites that are coupled to distant functional sites. However, identifying cryptic pockets remains difficult. Most known cryptic sites were only identified after the serendipitous discovery of a small molecule that binds and stabilizes the open form of the pocket [74, 241]. Experimental techniques for intentionally identifying and targeting cryptic pockets show great promise [77, 242, 243], but they still leverage the simultaneous discovery of ligands that bind and stabilize the open pocket. To overcome this limitation, a number of computational methods have been developed to identify cryptic pockets without requiring the simultaneous discovery of small molecules that bind them [47, 49, 244–249]. These methods have proved capable of retrodicting a number of previously identified cryptic pockets. More importantly, applications to a few established drug targets and other enzymes have successfully identified novel cryptic pockets that have been corroborated by subsequent experiments [49, 79, 250].

Here, we integrate atomically-detailed computer simulations and biophysical experiments to explore the potential utility of cryptic pockets in a target that has so far proved undruggable: the interferon inhibitory domain (IID) of Ebola viral protein 35 (VP35). Ebola virus causes a hemorrhagic fever that is often lethal, with case fatality rates approaching 90% in past outbreaks [251, 252]. While recent progress in vaccine development and use of biologics, such as antibodies, for therapeutic and prophylactic purposes show promise [252], small molecule drugs still offer many advantages, including ease of delivery, lower cost, and longer shelf life. The $\tilde{1}20$ residue IID of VP35 is a particularly appealing drug target for combating Ebola and other viruses in the Filoviridae family given that it has a well-conserved sequence and plays multiple essential roles in the viral lifecycle [253]. One of its primary functions is to antagonize a host's innate immunity, particularly RIG-I-like receptor (RLR)-mediated detection of

viral nucleic acids, to prevent an interferon (IFN) response and signaling of neighboring cells to heighten their antiviral defenses [254–256]. Crystal structures have provided a foundation for understanding much about the mechanism of VP35-mediated IFN antagonism [257, 258]. For example, they have revealed that VP35's IID binds both the blunt ends and backbone of double-stranded RNA (dsRNA), and that there is a PPI between these dsRNA-binding modes (Fig. 4.1) [258]. Disrupting any of these interactions could potentially render Ebola susceptible to a host's innate immunity. In particular, binding to dsRNA blunt ends plays a dominant role in IFN suppression by Ebola [259]. Indeed, mutations that reduce the IID's affinity for dsRNA blunt ends are sufficient to mitigate IFN antagonism, ultimately attenuating Ebola's pathogenicity [259–262]. So, disrupting this single binding mode could dramatically reduce the impact of an Ebola infection on the host and potentially reduce deleterious effects, including lethality. However, both dsRNA-binding interfaces are large flat surfaces that are difficult for small molecules to bind tightly (Fig. 4.1). As a result, [there are no approved therapies targeting VP35]. Both docking and screening attempts to discover small molecules that bind these interfaces have not yielded sufficiently strong leads to warrant clinical development [263, 264]. The discovery of cryptic pockets in VP35 could provide new opportunities for drugging this essential viral component.

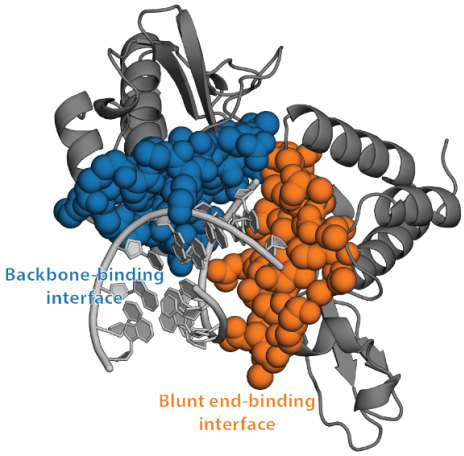


Figure 4.1: Crystal structure of two copies of VP35's IID (dark gray) bound to dsRNA (light gray) via two flat interfaces (PDB ID 3L25). The backbone-binding interface (blue) and blunt end-binding interface (orange) are shown as spheres to highlight that they lack deep pockets amenable to binding small molecules.

4.3 Results

4.3.1 Computer simulations reveal a potentially druggable cryptic pocket.

We applied our fluctuation amplification of specific traits (FAST) simulation algorithm [35] to enhance sampling of structures with large pocket volumes that may harbor cryptic pockets. FAST is a goal-oriented adaptive sampling algorithm that exploits Markov state model (MSM) methods to focus computational resources on exploring regions of conformational space with user-specified structural features. An MSM is a network model of a protein's energy landscape which consists of a set of structural states the protein adopts and the rates of hopping between them [53, 265]. Adaptive sampling algorithms enable efficient construction of MSMs by iteratively 1) running a batch of simulations, 2) building an MSM, and 3) selecting a subset of the states that have been identified so far as starting points for the next batch of simulations to maximize the chances of improving the model [266, 267]. FAST selects which states to further simulate in a manner that balances exploration/exploitation tradeoffs by considering 1) how well each state optimizes a user defined structural criterion (in this case maximizing the total pocket volume) and 2) the likelihood of discovering new conformational states [35]. After running FAST, we collected additional simulation data by launching each state on the Folding@home distributed computing environment, which brings together the computing resources of tens of thousands of citizen scientists who volunteer to run simulations on their personal computers. Our final model has 11,891 conformational states, providing a detailed characterization of the different structures the IID adopts but making manual interpretation of the model difficult.

To identify cryptic pockets within the large ensemble captured by our MSM, we applied our exposons analysis pipeline [49]. An exposon is a cluster of residues that undergo cooperative changes in their solvent exposure. Coupling between the solvent exposure of every pair of residues is quantified using a mutual information metric, as described in Methods. Exposons

are often associated with cryptic sites because the opening/closing of such pockets gives rise to cooperative increases/decreases in the solvent exposure of surrounding residues. Importantly, once an exposon has been identified, our MSM framework provides a facile means to identify the conformational changes that give rise to that exposon.

The IID has two significant exposons, one of which corresponds to a large cryptic pocket. The blue exposon (Fig. 4.2A and 4.2B) consists of a set of strongly-coupled residues in helix 7 and adjacent loops and secondary structure elements. Visualizing the conformational change that gives rise to this exposon reveals a substantial displacement of helix 7, creating a large cryptic pocket between it and the helical domain (Fig. 4.2C). A number of residues that are displaced along with helix 7 (i.e. A306, K309, and S310) make van der Waals contacts with the dsRNA backbone in the dsRNA-bound crystal structure [258], so targeting this cryptic pocket could directly disrupt this binding mode. Retrospective analysis of other validated drug targets suggests cryptic sites created by the movement of secondary structure elements, such as the displacement of helix 7, are often druggable [268]. The potential druggability of this cryptic site is also supported by application of the FTMap algorithm [269, 270], which predicts a number of hotspots within the pocket where small molecules could form a variety of energetically-favorable interactions (C.1). Unfortunately, disrupting backbone binding is of less therapeutic utility than disrupting blunt end binding and it is unknown whether the contacts between A306, K309, and S310 are essential for backbone binding. Therefore, it is unclear from this analysis alone whether drugging this newly discovered cryptic pocket would be useful.

The second exposon (orange in Fig. 4.2) encompasses portions of both dsRNA-binding interfaces, but it does not correspond to a cryptic pocket. This exposon includes residues that bind dsRNA's backbone (i.e. S272) and residues that interact with both the blunt ends and backbone of dsRNA (i.e. F239, Q274, and I340) [258]. Therefore, altering the conformational preferences of the second exposon could potentially disrupt the blunt end-binding mode and its crucial role in Ebola's ability to evade an immune response. However, the largest conformational change involved in the formation of this exposon is a displacement of the loop between

helices 3 and 4 (Fig. 4.2D). This rearrangement does not create a cryptic pocket that is large enough to accommodate drug-like molecules, so it is not obvious how to directly manipulate the orange exposon.

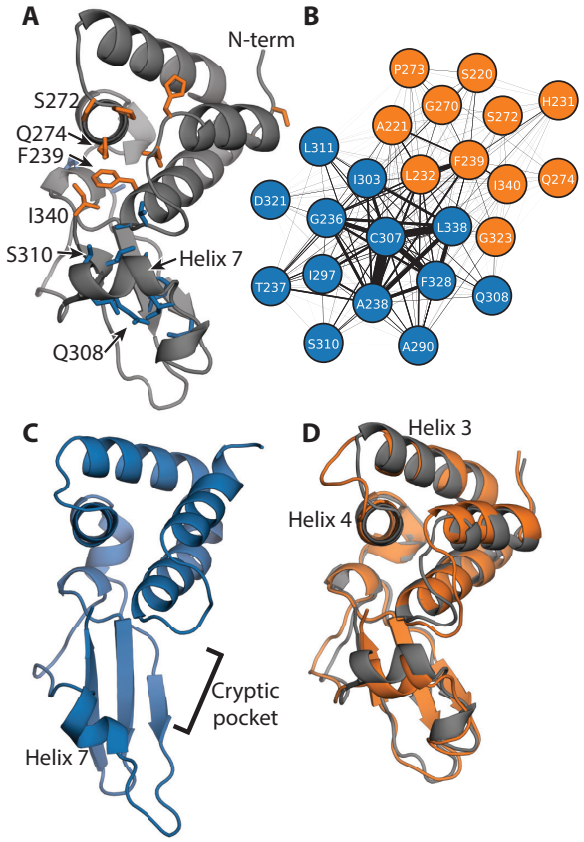


Figure 4.2: Exposons identify a large cryptic pocket and suggest potential allosteric coupling. **A.** Structure of VP35's IID highlighting residues in two exposons (blue and orange), the N-terminus (N-term), and C-terminus (I340). **B.** Network representation of the coupling between the solvent exposure of residues in the two exposons. The edge width between residues is proportional to the mutual information between them. **C.** Structure highlighting the opening of a cryptic pocket via the displacement of helix 7 that gives rise to the blue exposon. **D.** Structure highlighting the conformational change that gives rise to the orange exposon overlaid on the crystal structure (gray) to highlight that the rearrangements are subtler than in the blue exposon.

4.3.2 The cryptic pocket is allosterically coupled to the blunt end-binding interface.

Even though the cryptic pocket does not coincide with the interface of VP35's IID that binds dsRNA blunt ends, it could still serve as a cryptic allosteric site that allosterically controls RNA

binding. Indeed, the physical proximity of the two exposons and the coupling between them both hint at the possibility for allosteric coupling. Furthermore, our exposons analysis could easily underestimate this coupling given that it focuses on correlated transitions of residues between solvent exposed and completely buried states, leaving it blind to more subtle conformational fluctuations and the coupling of residues that are always buried (or always exposed).

To explore the potential for a broader allosteric network, we applied the correlation of all rotameric and dynamical states (CARDS) algorithm [90]. CARDS classifies each dihedral in each snapshot of a simulation as being in one of three rotameric states (gauche+, gauche-, or trans) and one of two dynamical states (ordered or disordered). A dihedral is said to be disordered if it is rapidly hopping between different structural states, and it is classified as ordered if it appears to be locked into a single rotameric state for a prolonged time. The mutual information metric is then used to quantify how strongly coupled the structural and dynamical states of each pair of dihedrals are, enabling CARDS to capture the roles of both concerted structural changes and conformational entropy in allosteric communication. Importantly, CARDS accounts for the potential role of residues that are always buried or always exposed to solvent and subtle conformational changes that do not alter the solvent exposure of residues.

CARDS reveals a broader allosteric network than that identified by our exposons analysis and suggests strong coupling between the cryptic pocket and blunt end-binding interface (Fig. 4.3). This network consists of five communities of strongly coupled residues, four of which coincide with large portions of the two dsRNA-binding interfaces. One of these communities (orange) is a hub in the network, having significant coupling to all the other communities. It encompasses part of the orange exposon, particularly residues around the loop between helices 3 and 4. The orange CARDS community and exposon both capture Q274, which engages in both dsRNA-binding interfaces, and S272, which contacts the backbone [258]. However, the CARDS community includes many additional residues not captured by exposons analysis. Examples include I278, which engages in both dsRNA-binding interfaces, and D271, which is part of the PPI between the two binding modes [258]. One of the orange community's strongest allosteric

connections is to the green community. This community encompasses the rest of the residues in the orange exposon, including F239 and I340, which are part of both dsRNA-binding interfaces [258]. The green community also captures additional residues, reaching deep into the helical domain. The orange community is also strongly coupled to the blue community, which includes much of helix 7 and nearby residues that move to give rise to the cryptic pocket that was captured by the blue exposon. Notably, the orange and blue communities are both coupled to a cyan cluster that was not hinted at by our exposons analysis because the residues involved are always solvent exposed. It includes R322, which is part of the blunt end-binding interface and the PPI between the two binding modes, and K282, which also contacts dsRNA blunt ends [258]. In addition, this community includes K339, which is an important determinant of the electrostatic favorability of dsRNA binding [258]. Together, these results suggest that opening of the cryptic pocket could strongly impact residues involved in both dsRNA-binding interfaces, as well as the PPI between the two binding modes.

To determine the relative importance of the structural and dynamical preferences of this community, we compared the magnitudes of the structural and dynamical components of CARDS. This analysis revealed that concerted structural changes are the dominant mode of allosteric communication in the IID, rather than conformational entropy and dynamical allostery (C.2). Therefore, examining structures where the orange community undergoes large conformational changes might reveal the perturbations these motions induce elsewhere in the protein.

To understand the potential impact of targeting the cryptic pocket on the blunt end-binding mode, we performed a dimensionality reduction based on the orange community. Since the orange community is a hub in the allosteric network, we reasoned that performing a dimensionality reduction based on the structural preferences of this community and examining representative structures would report on what is happening throughout the protein. To understand what sort of conformational changes are present, we performed a dimensionality reduction on our simulation dataset by applying principal component analysis (PCA) to the distances between the Cβ atoms of every pair of residues in the orange community. Projecting our MSM onto the

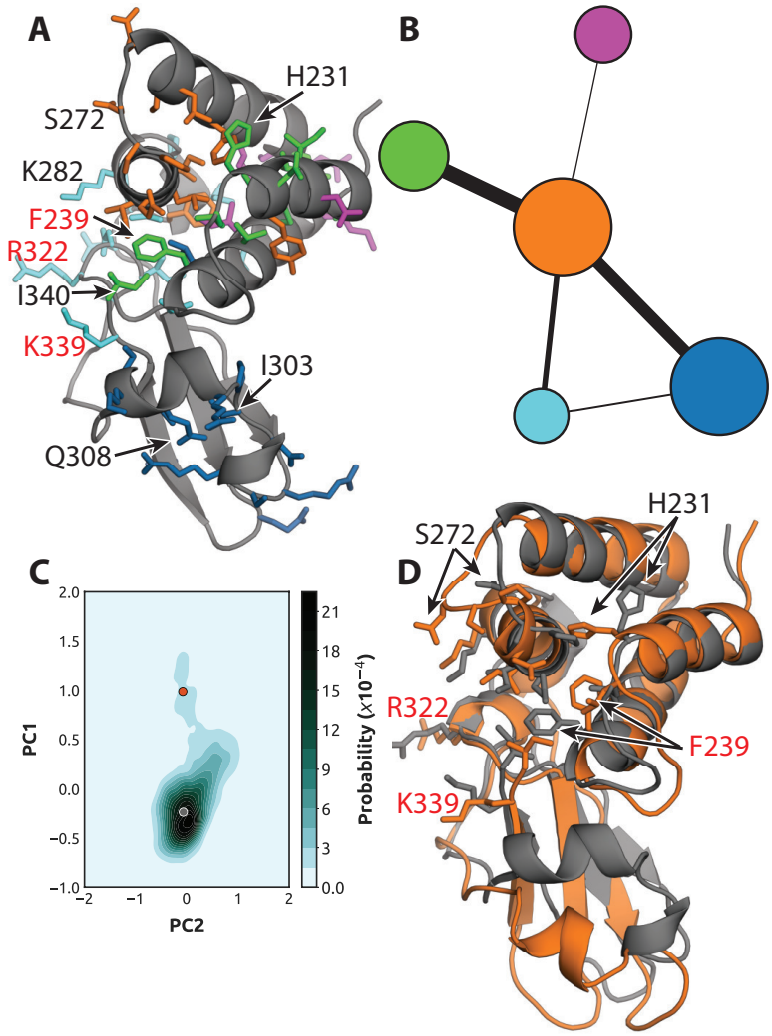


Figure 4.3: eVP35 allosteric network revealed by the CARDS algorithm. **A.** Structure of VP35's IID with residues in the allosteric network shown in sticks and colored according to which of five communities they belong to. Substitution of residues labeled in red with alanine disrupts binding to dsRNA blunt ends and results in a dramatic reduction in immune suppression. **B.** Network representation of the coupling between communities of residues, colored as in A. Node size is proportional to the strength of coupling between residues in the community, and edge widths are proportional to the strength of coupling between the communities. **C.** Free energy landscape of the orange exposon projected onto the first two principal components, PC1 and PC2, highlighting the centroid structures of the free energy minimum (gray circle) and excited state (orange circle). **D.** Structures of the centroids (colored as in panel C) capture opening of the cryptic pocket and rearrangements involving key residues for PPIs and PNIs.

first two principal components (PC1 and PC2) reveals one dominant free energy minimum and a broad excited state (Fig. 4.3C).

Comparing representative structures for the orange community's two dominant states suggests

the cryptic pocket is indeed a cryptic allosteric site, targeting of which could allosterically disrupt binding of VP35's IID to dsRNA blunt ends. Most importantly, conformational changes of the orange community are associated with opening of the cryptic pocket (Fig. 4.3D). Therefore, targeting the cryptic pocket could modulate the entire allosteric network in addition to its potential direct effect on the backbone-binding mode. Comparing the structures also reveals that the end of helix 4 frays and the preceding loop, which sits at the PPI between the two dsRNA-binding modes, is displaced. So, targeting the cryptic pocket could allosterically modulate this PPI. Finally, we note a substantial reshuffling of residues F239, H231, and P273 and modest displacements of R322 and K339. Previous work has demonstrated that F239A, R322A, and K339A substitutions are each sufficient to disrupt dsRNA binding and IFN suppression [258]. CARDS analysis suggests targeting the cryptic pocket could allosterically alter the structures of these residues and have a similar impact on dsRNA binding.

4.3.3 Thiol labeling experiments corroborate the predicted cryptic pocket.

One way to experimentally test our prediction of a cryptic pocket is to probe for solvent exposure of residues that are buried in available crystal structures but become exposed to solvent upon pocket opening. Cysteines are particularly appealing candidates for such experiments because 1) they have a low abundance and 2) their thiol groups are highly reactive, so it is straightforward to detect exposed cysteines by introducing labeling reagents that covalently bind accessible thiols. Fortuitously, VP35's IID has two cysteines (C307 and C326) that are buried in available crystal structures but become exposed to solvent when the cryptic pocket opens (Fig. 4.4A). There is also a cysteine (C275) that is on the surface of the apo crystal structure [257] and a fourth cysteine (C247) that is buried in the helical bundle. C275 is typically solvent exposed in our simulations, as expected based on the crystallographic data. Examining the solvent exposure of C247 revealed it is sometimes exposed to solvent via an opening of helix 1 relative to the rest of the helical bundle (C.3), but FTMap did not identify any hotspots that are likely to bind drug-like molecules in this region. Therefore, we expect to observe labeling

of all four cysteines on a timescale that is faster than global unfolding of the protein.

To experimentally test our predicted pocket, we applied a thiol labeling technique that probes the solvent exposure of cysteine residues [271]. For these experiments, 5,5'-Dithiobis-(2-Nitrobenzoic Acid) (also known as DTNB or Ellman's reagent, Fig. 4.4B) is added to a protein sample. Upon reaction with the thiol group of an exposed cysteine, DTNB breaks into two TNB molecules, one of which remains covalently bound to the cysteine while the other is released into solution. The accumulation of free TNB can be quantified based on the increased absorbance at 412 nm. We have previously applied this technique to test predicted pockets in β-lactamase enzymes [49, 147].

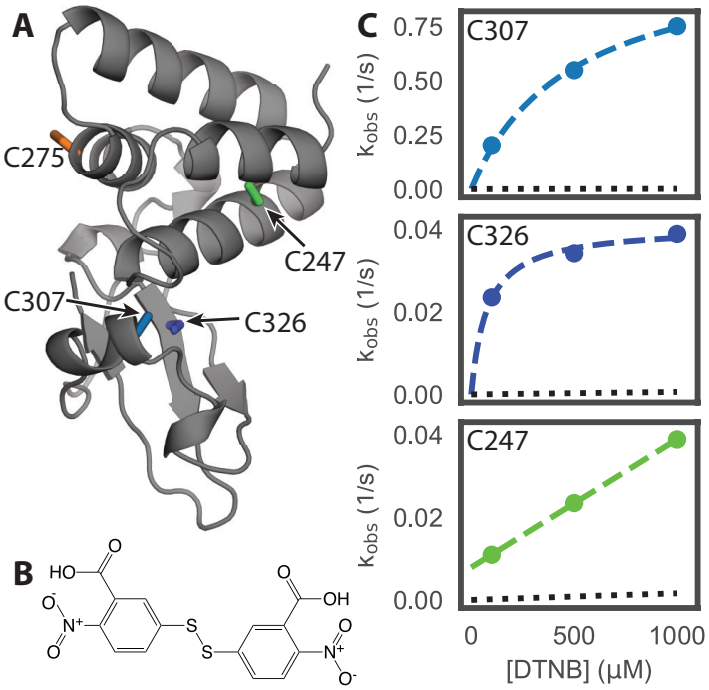


Figure 4.4: Thiol labeling supports the existence of the predicted cryptic pocket. Structure of VP35's IID highlighting the locations of the four native cysteines (sticks). C307 and C326 are both buried and point into the proposed cryptic pocket. B) Structure of the DTNB labeling reagent. C) Observed labeling rates (circles) at a range of DTNB concentrations. Fits to the Linderstrøm-Lang model are shown in dashed colored lines and the expected labeling rate from the unfolded state is shown as black dotted lines. The mean and standard deviation from three replicates are shown but error bars are generally smaller than the symbols. Labeling for C275 is not shown because it is surface exposed in both the available crystal structures and our simulations, and it behaves as expected (labeling rate greater than 1 s^{-1} with a linear dependence on [DTNB]).

As expected from our computational model, the observed signal from our thiol labeling experiments is consistent with opening of the cryptic pocket (Fig. 4.4C). Absorbance curves are best fit by four exponentials, each with an approximately equivalent amplitude that is consistent with expectations based on the extinction coefficient for DTNB (C.4). To assign these labeling rates to individual cysteines, we systematically mutated the cysteines to serines, performed thiol labeling experiments, and assessed which rates disappeared and which remained (C.6). For example, labeling of the C275S variant lacks the very fastest rate for wild-type, consistent with the intuition that a residue that is surfaced exposed in the crystal structure (i.e. C275) should label faster than residues that are generally buried. To test whether the observed labeling could be due to an alternative process, such as global unfolding, we determined the population of the unfolded state and unfolding rate of VP35's IID under native conditions (C.7) and the intrinsic labeling rate for each cysteine (C.6). As shown in Fig. 4.4C, the observed labeling rates are all considerably faster than the expected labeling rate from the unfolded state at a range of DTNB concentrations. This result confirms that labeling of all four cysteines arises from fluctuations within the native state, consistent with our computational predictions. Furthermore, the exposure of C247 is far rarer than C307 or C326 (equilibrium constants for the exposure of C247 and C307 are $5.4{\times}10^{-4}{\pm}8.1{\times}10^{-}6$ and $8.5{\times}10^{-2}{\pm}2.8{\times}10^{-3}$, respectively). Therefore, a ligand would have to pay a greater energetic cost to stabilize the conformational change that exposes C247 than to stabilize the open state of the cryptic allosteric site created by the motion of helix 7.

4.3.4 Stabilizing the open cryptic pocket allosterically disrupts binding to dsRNA blunt ends.

We reasoned that covalent attachment of TNB to C307 and C326 would provide a means to capture the open pocket and assess the impact of stabilizing this state on dsRNA binding. Addition of TNB to these cysteines is sterically incompatible with the closed conformation

of VP35's IID that has been observed crystallographically. TNB's mass of ~198 Da is also similar to many drug fragments used in screening campaigns, making it a reasonable surrogate for the type of effect one might achieve with a fragment hit. Given that we already know DTNB labels the IID's cysteines, a TNB-labeled sample is easily obtainable by waiting until the labeling reaction goes to completion. Finally, we have previously used this same strategy to identify cryptic pockets that exert allosteric control over the activity of β-lactamase enzymes [49, 147]. To ensure that we primarily capture the effect of labeling on pocket opening, we used a C247S/C275S variant of VP35's IID that only has cysteines pointing into the cryptic pocket. As with the wild-type protein, thiol labeling of the C247S/C275S variant is consistent with the formation of the proposed cryptic pocket (Supplementary Fig. C.5).

To measure the effect of TNB labeling on the IID's interaction with dsRNA, we developed a fluorescence polarization (FP) assay for monitoring dsRNA binding. Paralleling our past work on VP35-peptide interactions [272], we added varying concentrations of C247S/C275S IID to a fixed concentration of 25-bp dsRNA with a fluorescein isothiocyanate (FITC) conjugation at one end (C.7). Free FITC-dsRNA emits depolarized light upon excitation with polarized light because of the molecule's fast rotation. Binding of one or more VP35 molecules restricts the motion of FITC-dsRNA, resulting in greater emission of polarized light, which is best monitored by the change in anisotropy [273].

Monitoring the binding of unlabeled protein to 25-bp dsRNA with either blunt ends or 3' overhangs demonstrates that our FP assay is sensitive to both dsRNA-binding modes and gives affinities that are consistent with past work. Past work using a dot-blot assay to measure binding reported an apparent dissociation constant (K_d) for blunt-ended dsRNA of 3.4$\pm$0.07 $\pm$M [259]. Furthermore, sterically hindering binding of the IID to dsRNA blunt ends by adding 2-nucleotide overhangs to the 3' of the RNA reduces the apparent dsRNA-binding affinity by 10-fold [274]. This weaker interaction was attributed to the backbone-binding mode since it is still available to VP35's IID even when the presence of an overhang inhibits blunt end binding. Similarly, our FP assay gives an apparent K_d of 3.6$\pm$0.34 μM for blunt-ended dsRNA (Fig.

5A). Addition of 3' overhangs results in a strong rightwards shift of the binding curve, consistent with at least a 5-fold reduction in the apparent binding affinity (apparent K_d of 20.4±1.1 μM). However, an upper baseline could not be captured due to limitations in the protein's solubility, so this apparent Kd is a lower bound. The data are also fit well assuming an apparent Kd of 30.1±7.2 μM that was reported previously [274].

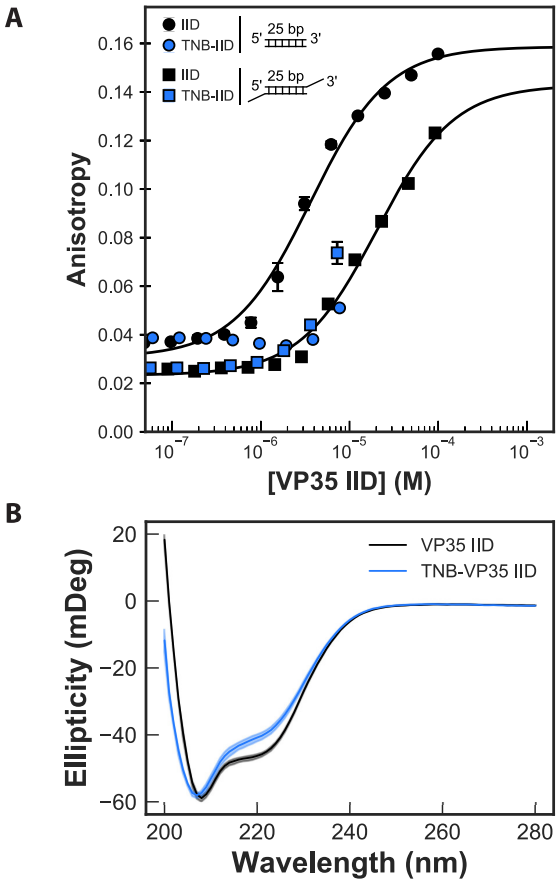


Figure 4.5: Stabilizing the open cryptic pocket in VP35's IID disrupts dsRNA binding. **A.** Binding of both TNB-labeled and unlabeled C247S/C275S variants of the IID to two different dsRNA constructs. This protein variant only has cysteines in the cryptic pocket. The RNA constructs both have a 25-bp double-stranded segment, and one has 2 nucleotide overhangs on the 3' ends. The anisotropy was measured via a fluorescence polarization assay and fit to a single-site binding model (black lines). The mean and standard deviation from three replicates are shown but error bars are generally smaller than the symbols. **B.** Circular dichroism (CD) spectra of labeled and unlabeled protein demonstrate that labeling does not unfold the protein. The opaque and semi-transparent lines represent the mean and standard deviation, respectively, from three replicates.

Repeating our FP assay with TNB-labeled protein reveals that labeling allosterically reduces the affinity for blunt-ended dsRNA by at least 5-fold (Fig. 4.5A). Solubility limitations again

prevented us from observing complete binding curves for labeled protein, but the data are sufficient to demonstrate that TNB-labeling has at least as strong an effect on binding as addition of a 3' overhang. As a control to ensure that labeling does not disrupt binding by simply unfolding the protein, we measured the circular dichroism (CD) spectra of labeled and unlabeled protein. The similarity between the CD spectra (Fig. 4.5B) demonstrates that the IID's overall fold is not grossly perturbed. The slight decrease in helicity at 220 nm can be attributed to the covalent modification disrupting the stability of helix 5 potentially causing local unfolding of this motif. Since the cryptic pocket does not coincide with the blunt end-binding interface, our results suggests the impact on dsRNA binding is allosteric. Furthermore, past work demonstrated that reducing the blunt end-binding affinity by as little as 3-fold is sufficient to allow a host to mount an effective immune response [258], so targeting our cryptic pocket could be of great therapeutic value.

4.4 Discussion.

We have identified a cryptic allosteric site in the IID of the Ebola VP35 protein that provides a new opportunity to target this essential viral component. Past work identified several sites within the VP35 IID that are critical for immune evasion and viral replication [253, 256, 261, 262], but structural snapshots captured crystallographically lacked druggable pockets [257, 258]. We used adaptive sampling simulations to access more of the ensemble of conformations that VP35 adopts, uncovering an unanticipated cryptic pocket. While the pocket directly coincides with the interface that binds the backbone of dsRNA, it was not clearly of therapeutic relevance since binding dsRNA's blunt ends is more important for Ebola's immune evasion mechanism [259]. However, our simulations also suggested the cryptic pocket is allosterically coupled to the blunt end-binding interface and, therefore, could modulate this biologically-important interaction. Subsequent experiments confirmed that fluctuations within the folded state of the IID expose two buried cysteines that line the proposed cryptic pocket

to solvent. Moreover, covalently modifying these cysteines to stabilize the open form of the cryptic pocket allosterically disrupts binding to dsRNA blunt ends by at least 5-fold. Previous work demonstrated that reducing the binding affinity by as little as 3-fold is sufficient to allow a host to mount an effective immune response [258]. Therefore, it may be possible to attenuate the impact of viral replication and restrict pathogenicity by designing small molecules to target the cryptic allosteric site we report here.

More generally, our results speak to the power of simulations to provide simultaneous access to both hidden conformations and dynamics with atomic resolution. Such information is extremely difficult to obtain from single structural snapshots or powerful techniques that report on dynamics without directly yielding structures, such as NMR and hydrogen deuterium exchange. As a result, simulations are a powerful means to uncover unanticipated features of proteins' conformational ensembles, such as cryptic pockets and allostery, providing a foundation for the design of further experiments. We anticipate such simulations will enable the discovery of cryptic pockets and cryptic allosteric sites in other proteins, particularly those that are currently considered undruggable. Furthermore, the detailed structural insight from simulations will facilitate the design of small molecule drugs that target these sites.

4.5 Methods

4.5.1 Molecular dynamics simulations and analysis

Simulations were initiated from chain B of PDB 3L25 [258] and run with Gromacs [145] using the amber03 force field [146] and TIP3P explicit solvent [143] at a temperature of 300 K and 1 bar pressure, as described previously [9]. We first applied our FAST-pockets algorithm [35] to balance 1) preferentially simulating structures with large pocket volumes that may harbor cryptic pockets with 2) broad exploration of conformational space. For FAST, we performed 10 rounds of simulations with 10 simulations/round and 80 ns/simulation. To acquire better

statistics across the landscape, we performed an RMSD-based clustering using a hybrid k-centers/k-medoids algorithm [230] implemented in Enspara [39] to divide the data into 1,000 clusters. Then we ran three simulations initiated from each cluster center on the Folding@home distributed computing environment, resulting in an aggregate simulation time of 122 μs.

Exposons were identified using our previously described protocols,11 as implemented in Enspara [39]. Briefly, the solvent accessible surface area (SASA) of each residue's side-chain was calculated using the Shrake-Rupley algorithm [275] implemented in MDTraj [148] using a drug-sized probe (2.8 Å sphere). Conformations were clustered based on the SASA of each residue using a hybrid k-centers/k-medoids algorithm, using a 2.5 Å2 distance cutoff and 5 rounds of k-medoids updates. A Markov time of 6 ns was selected based on the implied timescales test (C.8). The center of each cluster was taken as an exemplar of that conformational state, and residues were classified as exposed if their SASA exceeded 2.0 Å2 and buried otherwise. The mutual information between the burial/exposure of each pair of residues was then calculated based on the MSM (i.e. treating the centers as samples and weighting them by the equilibrium probability of the state they represent). Finally, exposons were identified by clustering the matrix of pairwise mutual information values using affinity propagation [276].

The CARDS algorithm [90] was applied to identify allosteric coupling using our established protocols [48], as implemented in Enspara [90]. Briefly, each dihedral angle in each snapshot of the simulations was assigned to one of three rotameric states (gauche+, gauche-, or trans) and one of two dynamical states (ordered or disordered). The total coupling between each pair of dihedrals X and Y was then calculated as

$$I_H(X_R, Y_R) = I(X_R, Y_D) + I(X_R, Y_D) + I(X_D, Y_R) + I(X_D, Y_D) \tag{4.1}$$

where I is the mutual information metric, X_R is the rotameric state of dihedral X, and X_D is the dynamical state of dihedral X. The term $I(X_R, Y_R)$ is the purely structural coupling, while the sum of the other three terms is referred to as the disorder-mediated coupling. The dihedral

level couplings were coarse-grained into residue-level coupling by summing the total coupling between all the relevant dihedrals. Communities of coupled residues were identified by clustering the residue-level matrix of total couplings using affinity propagation [276]. The constructed network was subsequently filtered to only retain significant edges [277].These algorithms are available at github.com/bowman-lab.

4.5.2 Protein expression and purification

All variants of VP35's IID were purified from the cytoplasm of E. coli BL21(DE3) Gold cells (Agilent Technologies). Variants were generated using the site directed mutagenesis method and confirmed by DNA sequencing. Transformed cells were grown at 37°C until OD 0.3 then grown at 18°C until induction at OD 0.6 with 1 mM IPTG (Gold Biotechnology, Olivette, MO). Cells were grown for 15 hours then centrifuged after which the pellet was resuspended in 20 mM Sodium Phosphate pH 8, 1 M sodium chloride, with 5.1 mM β-mercaptoethanol. Resuspended cells were subjected to sonication at 4°C followed by centrifugation. The supernatant was then subjected to Ni-NTA affinity, TEV digestion, cation exchange (BioRad UNOsphere Rapid S column), and size exclusion chromatography (BioRad Enrich SEC 70 column) into 10 mM Hepes pH 7, 150 mM NaCl, 1 mM MgCl2, 2 mM TCEP.

4.5.3 Thiol labeling

We monitored the change in absorbance over time of 5,5'-dithiobis-(2-nitrobenzoic acid) (DTNB, Ellman's reagent, Thermo Fisher Scientific). Various concentrations of DTNB were added to protein and change in absorbance was measured in either an SX-20 Stopped Flow instrument (Applied Photophysics, Leatherhead, UK), or an Agilent Cary60 UV-vis spectrophotometer at 412 nm until the reaction reached steady state ($\tilde{3}$00 s). Data were fit with a Linderstrøm-Lang model to extract the thermodynamics and/or kinetics of pocket opening, as described in detail

previously [49]. As a control, the equilibrium constant for folding and the unfolding rate were measured (C.1) and used to predict the expected labeling rate from the unfolded state. The equilibrium constant was inferred from a two-state fit to urea melts monitored by fluorescence and unfolding rates were inferred from single exponential fits to unfolding curves monitored by fluorescence after the addition of urea, as described previously [49,147,278]. Fluorescence data were collected using a Photon Technology International Quanta- Master 800 rapid excitation spectrofluorometer with Quantum Northwest Inc. TC-125 Peltier-controlled cuvette holder.

4.5.4 Fluorescence polarization binding assay

Binding affinities between variants of VP35's IID and dsRNA were measured using fluorescence polarization in 10 mM Hepes pH 7, 150 mM NaCl, 1 mM MgCl2. A 25 base pair FITC-dsRNA (Integrated DNA Technologies) substrate with and without a 2 nucleotide 3' overhang was included at 100 nM. The sample was equilibrated for one hour before data collection. Data were collected on a BioTek Synergy2 Multi-Mode Reader as polarization and were converted to anisotropy as described previously [273]. TNB-labeled samples were generated by allowing DTNB and VP35's IID to react for 3 minutes and then removing excess DTNB with a Zeba spin desalting columns (Thermo Fisher Scientific). A single-site binding model was sufficient to fit the data.

4.6 Acknowledgements

We are grateful to the citizen scientists who participate in Folding@home for volunteering to run simulations on their personal computers. This work was funded by NSF CAREER Award MCB-1552471 and NIH grant R01 GM124007 (Bowman), as well as NIH grants R01AI123926, P01AI120943, and R01AI143292 (Amarasinghe). GRB holds a Career Award at the Scientific Interface from the Burroughs Wellcome Fund and a Packard Fellowship for Science

and Engineering from The David & Lucile Packard Foundation. MAC was supported by the 5R25GM103757 IMSD program and SS was supported by a MilliporeSigma Fellowship. We thank Drs. Timothy M. Lohman and Alexander G. Kozlov for advice on FP assays.

Chapter 5

The SARS-CoV-2 nucleocapsid protein is dynamic, disordered, and phase separates with RNA.

This chapter is adapted from the following publication:

Cubuk, J., Alston, J.J., Incicco, J.J., Singh, S., Stuchell-Brereton, M.D., Ward, M.D., Zimmerman, M.I., Vithani, N., Griffith, D., Wagoner, J.A., Bowman, G.R., Hall, K.B., Soranno, A., Holehouse, A.S., The SARS-CoV-2 nucleocapsid protein is dynamic, disordered, and phase separates with RNA, Available on Biorxiv: https://doi.org/10.1101/2020.06.17.158121 [2]

In this work, my work in setting up, simulating, and generating seed conformations of the folded domains (and populations) led to the data presented in figures 5.1, 5.2, and 5.4.

5.1 Abstract

The SARS-CoV-2 nucleocapsid (N) protein is an abundant RNA binding protein critical for viral genome packaging, yet the molecular details that underlie this process are poorly understood. Here we combine single-molecule spectroscopy with all-atom simulations to uncover the molecular details that contribute to N protein function. N protein contains three dynamic disordered regions that house putative transiently-helical binding motifs. The two folded domains interact minimally such that full-length N protein is a flexible and multivalent RNA binding protein. N protein also undergoes liquid-liquid phase separation when mixed with RNA, and polymer theory predicts that the same multivalent interactions that drive phase separation also engender RNA compaction. We offer a simple symmetry-breaking model that provides a plausible route through which single-genome condensation preferentially occurs over phase separation, suggesting that phase separation offers a convenient macroscopic readout of a key nanoscopic interaction.

5.2 Introduction

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is an enveloped, positive-strand RNA virus that causes the disease COVID-19 (Coronavirus Disease-2019) [279]. While coronaviruses typically cause relatively mild respiratory diseases, COVID-19 is on course to kill half a million people in the first six months since its emergence in late 2019 [279–281]. Given the timeframe for vaccine development is on the order of months to years, alternative therapeutic approaches are sought to ameliorate viral morbidity and mortality [282].

A challenge in identifying candidate drugs is our relatively sparse understanding of the molecular details that underlie the function of SARS-CoV-2 proteins. As a result, there is a surge of biochemical and biophysical exploration of these proteins, with the ultimate goal of identifying proteins that are suitable targets for disruption, ideally with insight into the molecular details

of how disruption could be achieved [283, 284].

While much attention has been focused on the Spike (S) protein, many other SARS-CoV-2 proteins play equally critical roles in viral physiology, yet we know relatively little about their structural or biophysical properties [285–288]. Here we performed a high-resolution structural and biophysical characterization of the SARS-CoV-2 nucleocapsid (N) protein, the protein responsible for genome packaging [289–291]. A large fraction of N protein is predicted to be intrinsically disordered, which constitutes a major barrier to conventional structural characterization [290]. To overcome these limitations, we combined single-molecule spectroscopy with all-atom simulations to build a residue-by-residue description of all three disordered regions in the context of their folded domains The combination of single-molecule spectroscopy and simulations to reconstruct structural ensembles has been applied extensively to uncover key molecular details underlying disordered protein regions [292–297]. Our goal here is to provide biophysical and structural insights into the physical basis of N protein function.

In exploring the molecular properties of N protein, we discovered it undergoes phase separation with RNA, as was also reported recently [298–300]. Given N protein underlies viral packaging, we reasoned phase separation may in fact be an unavoidable epiphenomenon that reflects physical properties necessary to drive compaction of long RNA molecules. To explore this principle further, we developed a simple physical model, which suggested symmetry breaking through a small number of high-affinity binding sites can organize anisotropic multivalent interactions to drive single-polymer compaction, as opposed to multi-polymer phase separation. Irrespective of its physiological role, our results suggest that phase separation provides a macroscopic readout (visible droplets) of a nanoscopic process (protein:RNA and protein:protein interaction). In the context of SARS-CoV-2, those interactions are expected to be key for viral packaging, such that assays which monitor phase separation of N protein with RNA may offer a convenient route to identify compounds that will also attenuate viral assembly.

5.3 Results

Coronavirus nucleocapsid proteins are multi-domain RNA binding proteins that play a critical role in many aspects of the viral life cycle [291,301]. The SARS-CoV-2 N protein shares a number of sequence features with other nucleocapsid proteins from coronaviruses (Fig. D.1-D.5). Work on N protein from a range of model coronaviruses has shown that N protein undergoes both self-association, interaction with other proteins, and interaction with RNA, all in a highly multivalent manner.

The SARS-CoV-2 N protein can be divided into five domains; a predicted intrinsically disordered N-terminal domain (NTD), an RNA binding domain (RBD), a predicted disordered central linker (LINK), a dimerization domain, and a predicted disordered C-terminal domain (CTD) (Fig. 5.1). While SARS-CoV-2 is a novel coronavirus, decades of work on model coronaviruses (including SARS coronavirus) have revealed a number of features expected to hold true in the SARS-CoV-2 N protein. Notably, all five domains are predicted to bind RNA [303–309], and while the dimerization domain facilitates the formation of well-defined stoichiometric dimers, RNA-independent higher-order oligomerization is also expected to occur [308, 310–312]. Importantly, protein-protein and protein-RNA interaction sites have been mapped to all three disordered regions.

Despite recent structures of the RBD and dimerization domains from SARS-CoV-2, the solution-state conformational behavior of the full-length protein remains elusive [313–315]. Understanding N protein function necessitates a mechanistic understanding of the flexible predicted disordered regions and their interplay with the folded domains. A recent small-angle X-ray study shows good agreement with previous work on SARS, suggesting the LINK is relatively extended, but neither the structural basis for this extension nor the underlying dynamics are known [303, 316].

Here, we address these questions by probing three three full-length constructs of the N protein

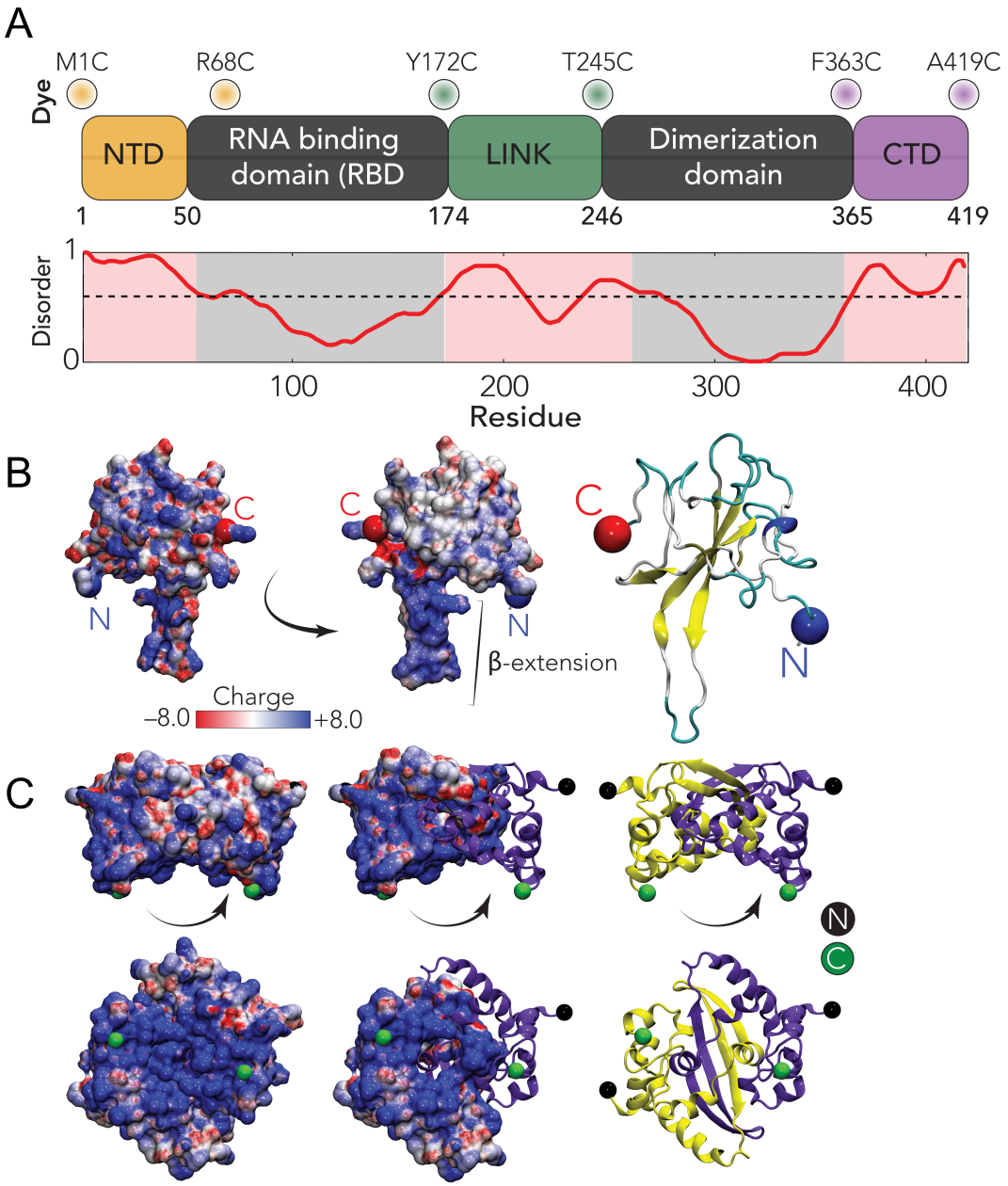


Figure 5.1: Sequence and structural summary of N protein . **A.** Domain architecture of the SARS-CoV-2 N protein. Dye positions used in this study are annotated across the top, disorder prediction calculated across the bottom. The specific positions were selected such that fluorophores are sufficiently close to be in the dynamic range of FRET measurements. Labeling was achieved using cysteine mutations and thiol-maleimide chemistry. **B.** Structure of the SARS-CoV-2 RNA binding domain (RBD) (PDB: 6yi3). Center and left: coloured based on surface potential calculated with the Adaptive Poisson Boltzmann Method [302], revealing the highly basic surface of the RBD. Right: ribbon structure with N- and C-termini highlighted. **C.** Dimer structure of the SARS-CoV-2 dimerization domain (PDB: 6yun). Center and left: coloured based on surface potential, revealing the highly basic surface. Right: ribbon structure with N- and C-termini highlighted.

with fluorescent labels (Alexa 488 and 594) flanking the NTD, the LINK, and the CTD (see Fig. 5.1A). These constructs allow us to probe conformations and dynamics of the disordered regions in the context of the full-length protein using single-molecule Förster Resonance Energy Transfer (FRET) and Fluorescence Correlation Spectroscopy (FCS) (see SI for details). In parallel to the experiments, we performed all-atom Monte Carlo simulations of each of the three IDRs in isolation and in context with their adjacent folded domains.

5.3.1 The NTD is disordered, flexible, and transiently interacts with the RBD.

We started our analysis by investigating the NTD conformations. Under native conditions, single-molecule FRET measurements revealed the occurrence of a single population with a mean transfer efficiency of 0.61 ± 0.03 (Fig. 5.2A and Fig. D.6). To assess whether this transfer efficiency reports about a rigid distance (e.g. structure formation or persistent interaction with the RBD) or is a dynamic average across multiple conformations, we first compare the lifetime of the fluorophores with transfer efficiency. Under native conditions, the donor and acceptor lifetimes for the NTD construct lie on the line that represents fast conformational dynamics (Fig. D.8A). To properly quantify the timescale associated with these fast structural rearrangements, we leveraged nanoseconds FCS. As expected for a dynamic population [317, 318], the cross-correlation of acceptor-donor photons for the NTD is anticorrelated (Fig. 5.2B and D.11). A global fit of the donor-donor, acceptor-acceptor, and acceptor-donor correlations yields a reconfiguration time τ_r = 170 $\pm$ 30 ns. This is longer than reconfiguration times observed for other proteins with a similar persistence length and charge content [318–321], hinting at a large contribution from internal friction due to rapid intramolecular contacts (formed either within the NTD or with the RBD) or transient formation of short structural motifs [322].

As a next step, we assessed the stability of the folded RBD and its influence on the conformations of the NTD by studying the effect of a chemical denaturant on the protein. The titration

with guanidinium chloride (GdmCl) reveals a decrease of transfer efficiencies when moving from native buffer conditions to 1 M GdmCl, followed by a plateau of the transfer efficiencies at concentrations between 1 M and 2 M and a subsequent further decrease at higher concentrations (Fig. D.6 and D.8). This behavior can be understood assuming that the plateau between 1 M and 2 M GdmCl represents the average of transfer efficiencies between two populations in equilibrium that have very close transfer efficiency and are not resolved because of shot noise. Indeed, this interpretation is supported by a broadening in the transfer efficiency peak between 1 M and 2 M GdmCl, which is expected if two overlapping populations react differently to denaturant. Besides the effect of the unfolding of the RBD, the dimensions of the NTD are also modulated by change in the solvent quality when adding denaturant (Fig. 5.2C, D.6, D.8) and this contribution to the expansion of the chain can be described using an empirical binding model [323–327]. A fit of the interdye root-mean-square distances to this model and the extracted stability of RBD (midpoint: 1.25 $\pm$ 0.2 M; ΔG_0 = (3 $\pm$ 0.6) RT) is presented in Fig. 5.2C A comparative fit of the histograms with two populations yields an identical result in terms of RBD stability and protein conformations (Fig. D.9).

These observations provide two important insights. Firstly, the RBD is completely folded under native conditions (Fig. 5.2C). Secondly, the RBD contributes significantly to the conformations of the measured NTD construct, mainly by reducing the accessible space of the disordered tail and favoring expanded configurations, as shown by the shift in transfer efficiency when the RBD is unfolded.

To better understand the sequence-dependent conformational behavior of the NTD we turned to all-atom simulations of an NTD-RBD construct. We used a novel sequential sampling approach that integrates long timescale MD simulations performed using the Folding@home distributed computing platform with all-atom Monte Carlo simulation performed with the ABSINTH forcefield to generate an ensemble of almost 400,000 distinct conformations (see methods [37, 328]. We also performed simulations of the NTD in isolation.

We observed excellent agreement between simulation and experiment for the equivalent inter-residue distance (Fig. 5.2D). The peaks on the left side of the histogram reflect specific simulations where the NTD engages more extensively with the RBD through a fuzzy interaction, leading to local kinetic traps [328]. We also identified several regions in the NTD where transient helices form, and using normalized distance maps found regions of transient attractive and repulsive interaction between the NTD and the RBD. In particular, the basic beta-strand extension from the RBD (Fig. 5.1B) repels the arginine-rich C-terminal region of the NTD, while a phenylalanine residue (F17) in the NTD engages with a hydrophobic face on the RBD (Fig. 5.2G). Finally, we noticed the arginine-rich C-terminal residues (residues 31 - 41) form a transient alpha helix projecting three of the four arginines in the same direction (Fig. 5.2H). These features provide molecular insight into previously reported functional observations (see Discussion).

5.3.2 The linker is highly dynamic and there is minimal interaction between the RBD and the dimerization domain.

We next turned to the linker (LINK) construct to investigate how the disordered region modulates the interaction and dynamics between the two folded domains. Under native conditions (50 mM Tris buffer), single-molecule FRET reveals a narrow population with mean transfer efficiency of 0.52 ± 0.03. Comparison of the fluorescence lifetime and transfer efficiency indicates that, like the NTD, the transfer efficiency represents a dynamic conformational ensemble sampled by the LINK (Fig. D.7B). ns-FCS confirms fast dynamics with a characteristic reconfiguration time τ_r of 120 ± 20 ns (Fig. 5.3B and D.11). This reconfiguration time is compatible with high internal friction effects, as observed for other unstructured proteins [318, 319], but may also account for the drag of the surrounding domains. The root-mean-square interdye distance for the LINK $r_{172-245}$ is equal to 57 ± 2 Å (l_p= 5.8 ± 0.4 Å) when assuming a Gaussian Chain distribution and 55 ± 2 Å (l_p= 5.4 ± 0.4 Å) when using a SAW model (see appendix E).

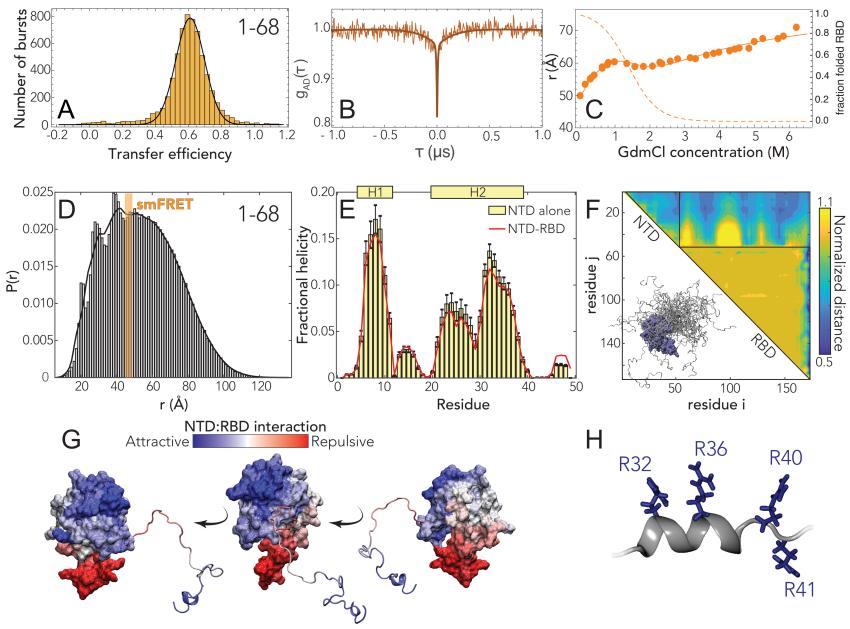


Figure 5.2: The N-terminal domain (NTD) is disordered with residual helical motifs. **A.** Histogram of the transfer efficiency distribution measured across the labeling positions 1 and 68 in the context of the full-length protein, under native conditions (50 mM Tris buffer). **B.** Donor-acceptor cross-correlation measured by ns-FCS (see Appendix E). The observed anticorrelated rise is the characteristic signature of FRET dynamics and the timescale associated is directly related to the reconfiguration time of the probed segment. **C.** Interdye distance as extracted from single-molecule FRET experiments across different concentrations using a Gaussian chain distribution, examining residues 1-68 in the context of the full length protein. The full line represents a fit to the model in Eq. D.19, which accounts for denaturant binding (see Table D.1) and unfolding of the folded RBD. The dashed line represents the estimate of folded RBD across different denaturant concentrations based on Eq. D.20 **D.** All-atom simulations of the NTD in the context of RBD reveal good agreement with smFRET-derived average distances. The peaks on the left shoulder of the histogram are due to persistent NTD-RBD interactions in a small subset of simulations. **E.** Transient helicity in the NTD in isolation or in the context of the RBD. Perfect profile overlap suggests interaction between the NTD and the RBD does not lead to a loss of helicity. **F.** Normalized distance maps (scaling maps) quantify heterogeneous interaction between every pair of residues in terms of average distance normalized by distance expected for the same system if the IDR had no attractive interactions (the "excluded volume" limit [329]). Both repulsive (yellow) and attractive (blue) regions are observed for NTD-RBD interactions. **G.** Projection of normalized distances onto the folded domain reveals repulsion is through electrostatic interaction (positively charged NTD is repelled by the positive face of the RBD, which is proposed to engage in RNA binding) while attractive interactions are between positive, aromatic, and polar residues in the NTD and a slightly negative and hydrophobic surface on the RBD (see Fig. 5.1B, center). **H.** The C-terminal half of transient helicity in H2 encodes an arginine-rich surface.

Next, we addressed whether the LINK segment populates elements of persistent secondary structure or forms stable interaction with the RBD or dimerization domains. Addition of the denaturant shows a continuous shift of the transfer efficiency toward lower values (Fig. D.6,D.8), that corresponds to an almost linear expansion of the chain (see Fig. 5.3C). These observations support a model in which LINK is unstructured and flexible and do not reveal a significant fraction of folding or persistent interactions with or between folded domains. Overall, our single-molecule observations report a relatively extended average inter-domain distance, suggesting a low number of interactions between folded domains. To further explore this conclusion, we turned again to Monte Carlo simulations.

As with the NTD, all-atom Monte Carlo simulations provide atomistic insight that can be compared with our spectroscopic results. Given the size of the system an alternative sampling strategy to the NTD-RBD construct was pursued here that did not include MD simulations of the folded domains, but we instead ran simulations of a construct that included the RBD, LINK and dimerization domain. In addition, we also performed simulations of the LINK in isolation.

We again found good agreement between simulations and experiment (Fig. 5.3D). The root mean square inter-residue distance between simulated positions 172 and 245 is 59.1 Å, which is within the experimental error of the single-molecule observations. Normalized distance map shows a number of regions of repulsion, notably that the RBD repels the N-terminal part of the LINK and the dimerization domain repels the C-terminal part of the LINK (Fig. 5.3E). We tentatively suggest this may reflect sequence properties chosen to prevent aberrant interactions between the LINK and the two folded domains. In the LINK-only simulations we identified two regions that form transient helices at low populations (20-25%), although these are much less prominent in the context of the full length protein (Fig. 5.3F). Those helices encompass a serine-arginine (SR) rich region known to mediate both protein-protein and protein-RNA interaction, and leads to the alignment of three arginine residues along one face of a helix. The second helix (H4) is a leucine/alanine-rich hydrophobic helix which may contribute to oligomerization, or act as a helical recognition motif for other protein interactions (notably as

a nuclear export signal for Crm1, see Discussion).

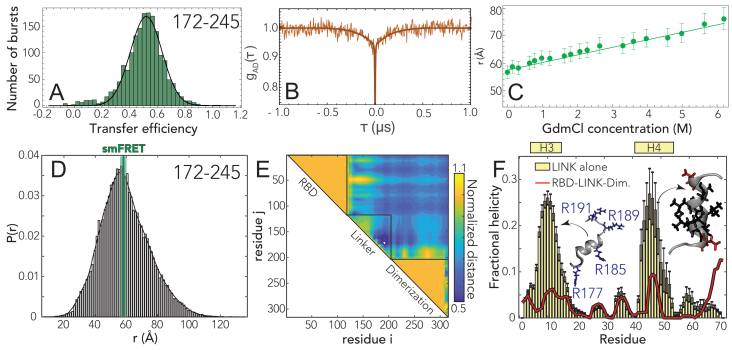


Figure 5.3: The RNA binding domain (RBD) and dimerization domains do not significantly interact and are connected by a disordered linker (LINK) **A.** Histogram of the transfer efficiency distribution measured across the labeling positions 172 and 245 in the context of the full-length protein, under native conditions (50 mM Tris buffer). **B.** Donor-acceptor cross-correlation measured by ns-FCS (see SI). The observed anticorrelated rise is the characteristic signature of FRET dynamics and the timescale associated is directly related to the reconfiguration time of the probed segment. **C.** Interdye distance as extracted from single-molecule FRET experiments across different denaturant concentrations. The full line represents a fit to the model in Eq. D.18, which accounts for denaturant binding. **D** Inter-residue distance distributions calculated from simulations (histogram) show good agreement with distances inferred from single-molecule FRET measurements (green bar). **E.** Scaling maps reveal repulsive interactions between the N- and C-terminal regions of the LINK with the adjacent folded domains. We also observe relatively extensive intra-LINK interactions around helix H4 (see Fig. 5.3F). **F.** Two transient helices are observed in the linker. The N-terminal helix H3 overlaps with part of the SR-region and orientates three arginine residues in the same direction, analogous to behavior observed for H2 in the NTD. The C-terminal helix H4 overlaps with a Leu/Ala rich motif which we believe is a conserved nuclear export signal (see Discussion).

5.3.3 The CTD engages in transient but non-negligible interactions with the dimerization domain.

Finally, we turned to the CTD. Single-molecule FRET experiments again reveal a single population with a mean transfer efficiency of 0.59 ± 0.03 (Fig. 5.4A) and the denaturant dependence follows the expected trend for a disordered region, with a shift of the transfer efficiency toward lower values (Fig. D.6,D.8), from 0.59 to 0.35. Interestingly, when studying the denaturant dependence of the protein, we noticed that the width of the distribution increases while moving toward native conditions. This suggests that the protein may form transient contacts or

adopt local structure. To investigate this aspect, we turned to the investigation of the dynamics. Though the comparison of the fluorophore lifetimes against transfer efficiency (Fig. D.7c) appears to support a dynamic nature underlying this population, nanosecond FCS reveals a flat acceptor-donor cross-correlation on the ns timescale (Fig. 5.4B). However, inspection of the donor-donor and acceptor-acceptor autocorrelations reveal a correlated decay with a characteristic time of 240 $\pm$ 50 ns. This is different from that expected for a completely static system such as polyprolines [330], where the donor-donor and acceptor-acceptor autocorrelation are also flat. An increase in the autocorrelations can be observed for static quenching of the dyes with aromatic residues.Interestingly, donor dye quenching can also contribute to a positive amplitude in the donor-acceptor correlation [331, 332]. Therefore, a plausible interpretation of the flat cross-correlation data is that we are observing two populations in equilibrium whose correlations (one anticorrelated, reflecting conformational dynamics, and one correlated, reflecting quenching due contact formation) compensate each other.

To further investigate the possible coexistence of these different species, we performed ns-FCS at 0.2 M GdmCl, where the width of the FRET population starts decreasing and the mean transfer efficiency is slightly shifted to larger values, under the assumption that the decreased width of the population reflects reduced interactions. Indeed, the cross-correlation of ns-FCS reveals a dynamic behavior with a reconfiguration time τ_r= 70 $\pm$ 15 ns (Fig. D.11). Based on these observations, we suggest that a very similar disordered population to the one observed at 0.2 M is also present under native conditions, but in equilibrium with a quenched species that forms long-lived contacts. Under the assumption that the mean transfer efficiency still originates (at least partially) from a dynamic distribution, the estimate of the inter-residue root-mean-square distance is $r_{363-419}$ = 51 $\pm$ 2 Å (l_p= 6.1 $\pm$ 0.4 Å) for a Gaussian chain distribution and $r_{363-419}$ = 49 $\pm$ 2 Å (l_p= 5.6 $\pm$ 0.4 Å) for the SAW model (see SI). However, some caution should be used when interpreting these numbers since we know there is some contribution from fluorophore quenching, which may in turn contribute to an underestimate of the effective transfer efficiency [333].

We again obtained good agreement between all-atom Monte Carlo simulations and experiment (Fig. 5.4D). We identified two transient helices, one (H5) is minimally populated but the second (H6) is more highly populated in the IDR-only simulation and still present at $\sim$20% in the folded state simulations (Fig. 5.4E). The difference reflects the fact that several of the helix-forming residues interact with the dimerization domain, leading to a competition between helix formation and intramolecular interaction. Scaling maps reveal extensive intramolecular interaction by the residues that make up H6, both in terms of local intra-IDR interactions and interaction with the dimerization domain (Fig. 5.4F). Mapping normalized distances onto the folded structure reveals that interactions occur primarily with the N-terminal portion of the dimerization domain (Fig. 5.4G). As with the LINK and the NTD, a positively charged set of residues immediately adjacent to the folded domain in the CTD drive repulsion between this region and the dimerization domain. H6 is the most robust helix observed across all three IDRs, and is a perfect amphipathic helix with a hydrophobic surface on one side and charged/polar residues on the other (Fig. 5.4H). The cluster of hydrophobic residues in H6 engage in intramolecular contacts and offer a likely physical explanation for the complex lifetime data.

5.3.4 N protein undergoes phase separation with RNA.

Over the last decade, biomolecular condensates formed through phase separation have emerged as a new mode of cellular organization [334–337]. Given the high interaction valency and the presence of molecular features similar to other proteins we had previously studied, we anticipated that N protein would undergo phase separation with RNA [338–340].

In line with this expectation, we observed robust droplet formation with homopolymeric RNA (Fig. 5.5A-B) under native buffer conditions (50 mM Tris) and at higher salt concentration (50 mM NaCl). Turbidity assays at different concentrations of protein and poly(rU) (200-250 nucleotides) demonstrate the classical reentrant phase behavior expected for a system undergoing heterotypic interaction (Fig. 5.5C-D). It is to be noted that turbidity experiments do not exhaus-

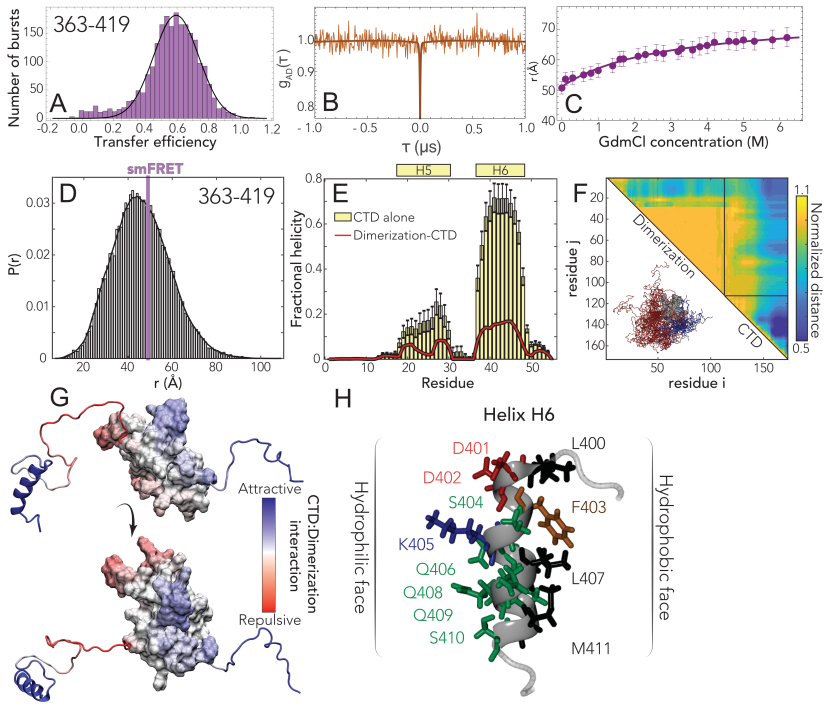


Figure 5.4: The C-terminal domain (CTD) is disordered, engages in transient interaction with the dimerization domain, and contains a putative helical binding motif. **A.** Histogram of the transfer efficiency distribution measured across the labeling positions 363 and 419 in the context of the full-length protein, under native conditions (50 mM Tris buffer). **B.** Donor-acceptor cross-correlation measured by ns-FCS (see Appendix E). The flat correlation indicates a lack of dynamics in the studied timescale or the coexistence of two populations in equilibrium whose correlations (one correlated and the other anticorrelated) compensate each other. **C.** Interdye distance as extracted from single-molecule FRET experiments across different denaturant concentrations. The full line represents a fit to the model in Eq. D.18, which accounts for denaturant binding. **D.** Inter-residue distance distributions calculated from simulations (histogram) show good agreement with distances inferred from single-molecule FRET measurements (purple bar). **E.** Two transient helices (H5 and H6) are observed in the CTD. Both show a reduction in population in the presence of the dimerization domain at least in part because the same sets of residues engage in transient interactions with the dimerization domain. **F.** Normalized contacts maps describe the average inter-residue distance between each pair of residues, normalized by the distance expected if the CTD behaved as a self-avoiding random coil. H6 engages in extensive intra-CTD interactions and also interacts with the dimerization domain. We observe repulsion between the dimerization domain and the N-terminal region of the CTD. **G.** The normalized distances are projected onto the surface to map CTD-dimerization interaction. The helical region drives intra-molecular interaction, predominantly with the N-terminal side of the dimerization domain. **H.** Helix H6 is an amphipathic helix with a polar/charged surface (left) and a hydrophobic surface (right).

tively cover all the conditions for phase separation and are only indicative of the low-boundary concentration regime explored in the current experiments. In particular, turbidity experiments

do not provide a measurement of tie-lines, though they are inherently a reflection of the free energy and chemical potential of the solution mixture [341]. Interestingly, phase separation occurs at relatively low concentrations, in the low μM range, which are compatible with physiological concentration of the protein and nucleic acids. Though increasing salt concentration results in an upshift of the phase boundaries, one has to consider that in a cellular environment this effect might be counteracted by cellular crowding.

One peculiar characteristic of our measured phase-diagram is the narrow regime of conditions in which we observe phase separation of nonspecific RNA at a fixed concentration of protein. This leads us to hypothesize that the protein may have evolved to maintain a tight control of concentrations at which phase separation can (or cannot) occur. Interestingly, when rescaling the turbidity curves as a ratio between protein and RNA, we find all the curve maxima aligning at a similar stoichiometry, approximately 20 nucleotides per protein in absence of added salt and 30 nucleotides when adding 50 mM NaCl (Fig. D.12). These ratios are in line with the charge neutralization criterion proposed by Banerjee et al., since the estimated net charge of the protein at pH 7.4 is +24 [342]. Finally, given we observed phase separation with poly(rU), the behavior we are observing is likely driven by relatively nonspecific protein:RNA interactions. In agreement, work from the Gladfelter [298], Fawzi [299], Zweckstetter [300], and Yildiz (unpublished) labs have also established this phenomenon across a range of solution conditions and RNA types.

Having established phase separation through a number of assays, we wondered what -if any- physiological relevance this may have for the normal biology of SARS-CoV-2.

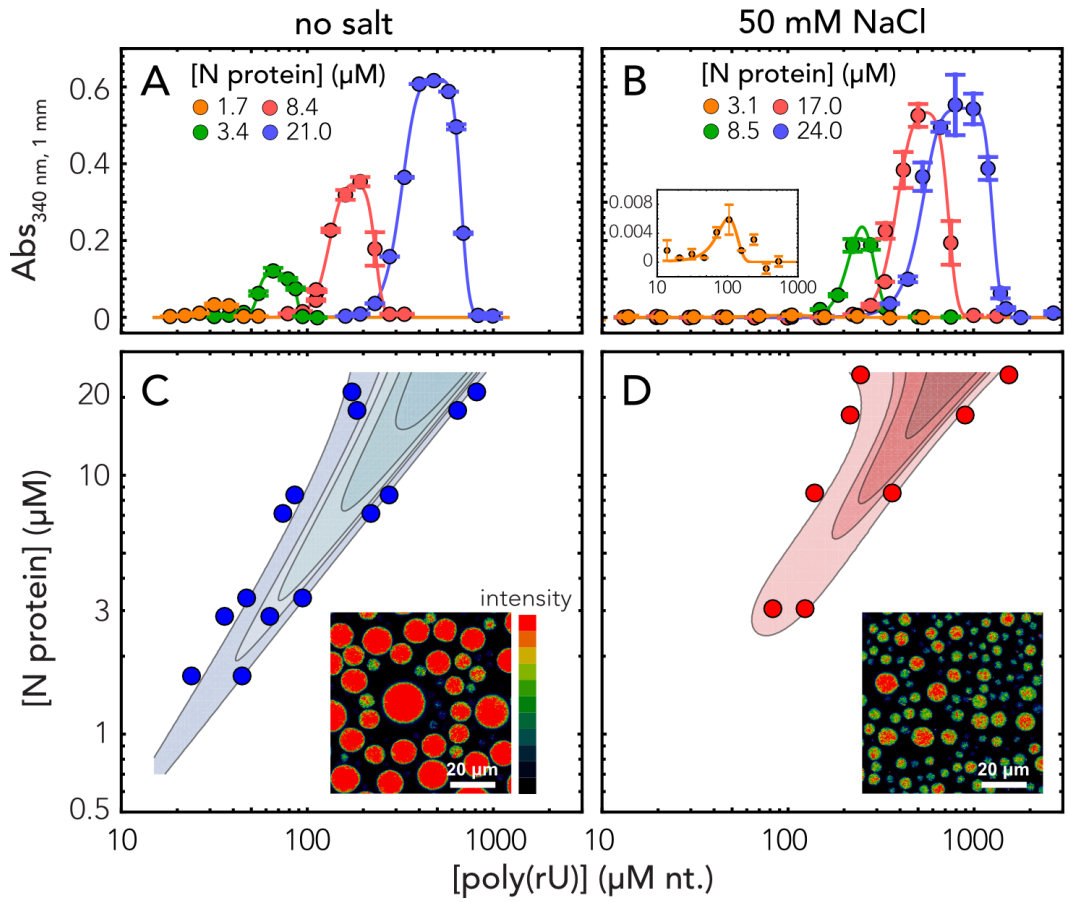


Figure 5.5: Nucleocapsid protein undergoes phase separation with RNA. **A-B.** Appearance of solution turbidity upon mixing was monitored to determine the concentration regime in which N protein and poly(rU) undergo phase separation. Representative turbidity titrations with poly(rU) in 50 mM Tris, pH 7.5 (HCl) at room temperature, in absence of added salt (A) and in presence of 50 mM NaCl (B), at the indicated concentrations of N protein. Points and error bars represent the mean and standard deviation of 2-4 consecutive measurements from the same sample. Solid lines are simulations of an empirical equation fitted individually to each titration curve (see SI). An inset is provided for the titration at 3.1 μM N protein in 50 mM NaCl to show the small yet detectable change in turbidity on a different scale. **C-D.** Projection of phase boundaries for poly(rU) and N protein mixtures highlights a re-entrant behavior, as expected for phase-separations induced by heterotypic interactions. Turbidity contour lines are computed from a global fit of all titrations curves (see Appendix E)

5.3.5 A simple polymer model shows symmetry-breaking can facilitate multiple metastable single-polymer condensates instead of a single multi-polymer condensate.

Why might phase separation of N protein with RNA be advantageous to SARS-CoV-2? One possible model is that large, micron-sized cytoplasmic condensates of N protein with RNA form through phase separation and play a role in genome packaging. These condensates may act as molecular factories that help concentrate the components for pre-capsid assembly (where we define a pre-capsid here simply as a species that contains a single copy of the genome with multiple copies of the associated N protein), a model that has been proposed in other viruses [343].

However, given that phase separation is unavoidable when high concentrations of multivalent species are combined, we propose that an alternative interpretation of our data is that in this context, phase separation is simply an inevitable epiphenomenon that reflects the inherent multi-valency of the N protein for itself and for RNA. This poses questions about the origin of specificity for viral genomic RNA (gRNA), and, of focus in our study, how phase separation might relate to a single genome packaging through RNA compaction.

Given the expectation of a single genome per virion, we reasoned SARS-CoV-2 may have evolved a mechanism to limit phase separation with gRNA (i.e. to avoid multi-genome condensates), with a preference instead for single-genome packaging (single-genome condensates). This mechanism may exist in competition with the intrinsic phase separation of the N protein with other nonspecific RNAs (nsRNA).

One possible way to limit phase separation between two components (e.g. gRNA/nsRNA and N protein) is to ensure the levels of these components are held at a sufficiently low total concentration such that the phase boundary is never crossed. While possible, suc,h a regulatory mechanism is at the mercy of extrinsic factors that may substantially modulate the saturation

concentration [**?**, 344–346]. Furthermore, not only must phase separation be prevented, but gRNA compaction should also be promoted through the binding of N protein. In this scenario, the affinity between gRNA and N protein plays a central role in determining the required concentration for condensation of the macromolecule (gRNA) by the ligand (N protein).

Given a defined valence of the system components, phase boundaries are encoded by the strength of interaction between the interacting domains in the components. Considering a long polymer (e.g. gRNA) with proteins adsorbed onto that polymer as adhesive points ("stickers"), the physics of associative polymers predicts that the same interactions that cause phase separation will also control the condensation of individual long polymers [338, 347–351]. With this in mind, we hypothesized that phase separation is reporting on the physical interactions that underlie genome compaction.

To explore this hypothesis, we developed a simple computational model where the interplay between compaction and phase separation could be explored. Our setup consists of two types of species: long multivalent polymers and short multivalent binders (Fig. 5.6A). All interactions are isotropic and each bead is inherently multivalent as a result. In the simplest instantiation of this model, favourable polymer:binder and binder:binder interactions are encoded, mimicking the scenario in which a binder (e.g. a protein) can engage in nonspecific polymer (RNA) interaction as well as binder-binder (protein-protein) interaction.

Simulations of binder and polymer undergo phase separation in a concentration-dependent manner, as expected (Fig. 5.6B,C). Phase separation gives rise to a single large spherical cluster with multiple polymers and binders (Fig. 5.6D, 5.6H). For a homopolymer, the balance of chain-compaction and phase separation is determined in part through chain length and binder K_d. In our system the polymer is largely unbound in the one-phase regime (suggesting the concentration of ligand in the one-phase space is below the K_d) but entirely coated in the two-phase regime, consistent with highly-cooperative binding behavior. In the limit of long, multivalent polymers with multivalent binders, the sharpness of the coil-to-globule transition is

such that an effective two-state description of the chain emerges, in which the chain is either expanded (non-phase separation-competent) OR compact (coated with binders, phase separation competent).

In light of these observations, we wondered if a break in the symmetry between intra- and intermolecular interactions would be enough to promote single-polymer condensation in the same concentration regime over which we had previously observed phase separation. Symmetry breaking in our model is achieved through a single high-affinity binding site (Fig. 5.6A). We choose this particular mode of symmetry-breaking to mimic the presence of a packaging signal -a region of the genome that is essential for efficient viral packaging- an established feature in many viruses (including coronaviruses) although we emphasize this is a general model, as opposed to trying to directly model gRNA with a packaging signal [352–354].

We performed identical simulations to those in Fig. 5.6C-D using the same system with polymers that now possess a single high affinity binding site (Fig. 5.6E). Under these conditions we did not observe large phase separated droplets (Fig. 5.6F). Instead, each individual polymer undergoes collapse to form a single-polymer condensate (Fig. 5.6E). Collapse is driven by the recruitment of binders to the high-affinity site, where they "coat" the chain, forming a local cluster of binders on the polymer. This cluster is then able to interact with the remaining regions of the polymer through weak "nonspecific" interactions, the same interactions that drove phase separation in Fig. 5.6B,C,D. Symmetry breaking is achieved because the local concentration of binder around the site is high, such that intramolecular interactions are favoured over intermolecular interaction. This high local concentration also drives compaction at low binder concentrations. As a result, instead of a single multi-polymer condensate, we observe multiple single-polymers condensates, where the absolute number matches the number of polymers in the system (Fig. 5.6G).

Our results can also be cast in terms of two distinct concentration (phase) boundaries - one for binder:high affinity site interaction (c_1), and a second boundary for "nonspecific" binder:polymer

interactions (c_2) at a higher concentration. c_2 reflects the boundary observed in Fig. 5.6C that delineated the one and two-phase regimes. At global concentrations below c_2, (but above c_1) the clustering of binders at a high affinity site raises the apparent local concentration of binders above c_2, from the perspective of other beads on the chain. In this way, a local high affinity binding site can drive "local" phase separation of a single polymer.

The high affinity binding site polarizes the single-polymer condensate, such that they are organized, recalcitrant to fusion, and kinetically stable. A convenient physical analogy is that of a micelle, which are non-stoichiometric stable assemblies. Even for micelles that are far from their optimal size, fusion is slow because it requires substantial molecular reorganization and the breaking of stable interactions [355, 356].

Finally, we ran simulations under conditions in which binder:polymer interactions were reduced, mimicking the scenario in which non-specific protein:RNA interactions are inhibited (Fig. 5.6L). Under these conditions no phase separation occurs for polymers that lack a high-affinity binding site, while for polymers with a high-affinity binding site no chain compaction occurs (in contrast to when binder:polymer interactions are present, see Fig. 5.6J). This result illustrates how phase separation offers a convenient readout for molecular interactions that might otherwise be challenging to measure.

We emphasize that our conclusions from simulations are subject to the parameters in our model. We present these results to demonstrate an example of "how this single-genome packaging could be achieved", as opposed to the much stronger statement of proposing "this is how it is" achieved. Recent elegant work by Ranganathan and Shakhnovich identified kinetically arrested microclusters, where slow kinetics result from the saturation of stickers within those clusters [357]. This is completely analogous to our results (albeit with homotypic interactions, rather than heterotypic interactions), giving us confidence that the physical principles uncovered are robust and, we tentatively suggest, quite general. Future simulations are required to systematically explore the details of the relevant parameter space in our system. However,

regardless of those parameters, our model does establish that if weak multivalent interactions underlie the formation of large multi-polymer droplets, those same interactions cannot also drive polymer compaction inside the droplet

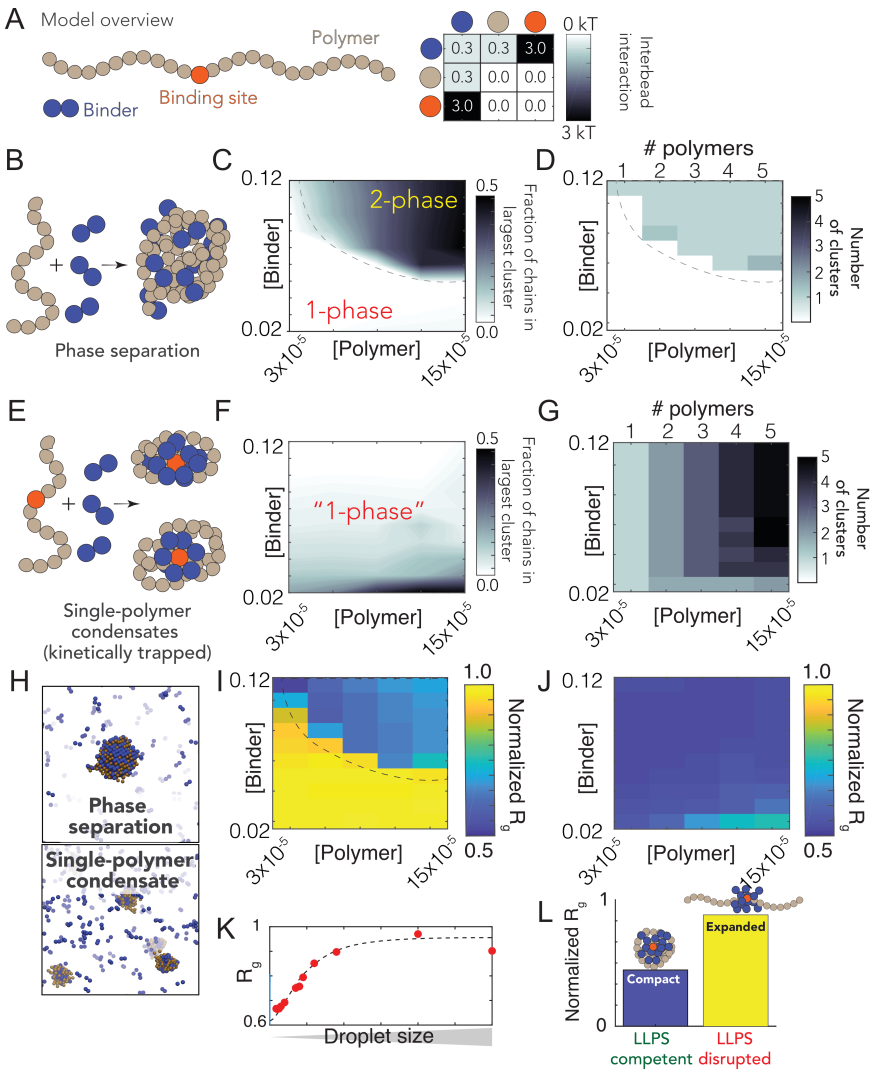


Figure 5.6: A simple polymer suggests symmetry breaking can promote single-polymer condensates over multi-polymer assemblies. **A.** Summary of our model setup, which involves long 'polymers' (61 beads) or short 'binders' (2 beads). Each bead is multivalent and can interact with every adjacent lattice site. The interaction matrix to the right defines the pairwise interaction energies associated with each of the bead times. **B.** Concentration dependent assembly behavior for polymers lacking a high-affinity binding site. **C.** Phase diagram showing the concentration-dependent phase regime - dashed line represents the binodal (phase boundary) and is provided to guide the eye. **D.** Analysis in the same 2D space as C assessing the number of droplets at a given concentration. When phase separation occurs a single droplet appears in almost all cases. **E.** TConcentration dependent assembly behavior for polymers with a high-affinity binding site. **F.** No large droplets are formed in any of the systems, although multiple polymer:binder complexes form. **G.** The number of clusters observed matches the number of polymers in the system - i.e. each polymer forms an individual cluster. **H.** Simulation snapshots from equivalent simulations for polymers with (top) or without (bottom) a single high-affinity binding site. **I.** Polymer dimensions in the dense and dilute phase (for the parameters in our model) for polymers with no high-affinity binding site. Note that compaction in the dense phase reflects finite-size effects, as addressed in panel K, and is an artefact of the relatively small droplets formed in our systems (relative to the size of the polymer). The droplets act as a bounding cage for the polymer, driving their compaction indirectly. **J.** Polymer dimensions across the same concentration space for polymers with a single high-affinity binding site. Across all concentrations, each individual polymer is highly compact. **K.** Compaction in the dense phase (panel I) is due to small droplets. When droplets are sufficiently large we observe chain expansion, as expected from standard theoretical descriptions. **L.** Simulations performed under conditions in which nonspecific interactions between binder and polymer are reduced (interaction strength = 0 kT). Under these conditions phase separation is suppressed.

5.4 Discussion

The nucleocapsid (N) protein from SARS-CoV-2 is a multivalent RNA binding protein critical for viral replication and genome packaging [289, 291]. To better understand how the various folded and disordered domains interact with one another, we applied single-molecule spectroscopy and all-atom simulations to perform a detailed biophysical dissection of the protein, uncovering several putative interaction motifs. Furthermore, based on both sequence analysis and our single-molecule experiments, we anticipated that N protein would undergo phase separation with RNA. In agreement with this prediction, and in line with work from the Gladfelter and Yildiz groups working independently from us, we find that N protein robustly undergoes phase separation in vitro with model RNA under a range of different salt conditions. Using simple polymer models, we propose that the same interactions that drive phase separation may also drive genome packaging into a dynamic, single-genome condensate. The formation of single-genome condensates (as opposed to multi-genome droplets) is influenced by the presence of one (or more) symmetry-breaking interaction sites, which we tentatively suggest could reflect packaging signals in viral genomes.

5.4.1 All three IDRs are highly dynamic

Our single-molecule experiments and all-atom simulations are in good agreement with one another, and reveal that all three IDRs are extended and highly dynamic. Simulations suggest the NTD may interact transiently with the RBD, which offers an explanation for the slightly slowed reconfiguration time measured by nanosecond FCS. The LINK shows rapid rearrangement, demonstrating the RBD and dimerization domain are not interacting. Finally, we see more pronounced interaction between the CTD and the dimerization domain, although these interactions are still highly transient.

Single-molecule experiments and all-atom simulations were performed on monomeric versions

of the protein, yet N protein has previously been shown to undergo dimerization and form higher-order oligomers in the absence of RNA [310]. To assess the formation of oligomeric species, we use a combination of nativePAGE, crosslinking and FCS experiments (see Fig. D.13). These experiments also verified that under the conditions used for single-molecule experiments the protein exists only as a monomer.

5.4.2 Simulations identify multiple transient helices

We identified a number of transient helical motifs which provide structural insight into previously characterized molecular interactions. Transient helices are ubiquitous in viral disordered regions and have been shown to underlie molecular interactions in a range of systems [343, 358–360].

Transient helix H2 (in the NTD) and H3 (in the LINK) flank the RBD and organize a set of arginine residues to face the same direction (Fig. 5.2E). Both the NTD and LINK have been shown to drive RNA binding, such that we propose these helical arginine-rich motifs (ARMs) may engage in both both nonspecific binding and may also contribute to RNA specificity, as has been proposed previously [303, 361, 362]. The serine-arginine SR-region (which includes H3) has been previously identified as engaging in interaction with a structured acidic helix in Nsp3 in the model coronavirus MHV, consistent with an electrostatic helical interaction [363, 364]. Recent NMR data also shows excellent agreement with our results, identifying a transient helix that shows 1:1 overlap with H3 [300].The SR-region is necessary for recruitment to replication-transcription centers (RTCs) in MHV, and also undergoes phosphorylation, setting the stage for a complex regulatory system awaiting exploration [365, 366].

Transient helix H4 (Fig. 5.3H), was previously predicted bioinformatically and identified as a conserved feature across different coronaviruses [303]. Furthermore, the equivalent region was identified in SARS coronavirus as a nuclear export signal (NES), such that we suspect this too is a classical Crm1-binding leucine-rich NES [367].

Transient helix H6 is an amphipathic helix with a highly hydrophobic face (Fig. 5.4H). Recent hydrogen-deuterium exchange mass spectrometry also identified H6 [315]. Residues in this region have previously been identified as mediating M-protein binding in other coronaviruses, such that we propose H6 underlies that interaction [368–370]. Recent work has also identified amphipathic transient helices in disordered proteins as interacting directly with membranes, such that an additional (albeit entirely speculative) role could involve direct membrane interaction, as has been observed in other viral phosphoproteins [371, 372].

5.4.3 The physiological relevance of nucleocapsid protein phase separation in SARS-CoV-2 physiology

Our work has revealed that SARS-CoV-2 N protein undergoes phase separation with RNA when reconstituted in vitro. The solution environment and types of RNA used in our experiments are very different from the cytoplasm and viral RNA. However, similar results have been obtained in published and unpublished work by several other groups under a variety of conditions, including via in cell experiments (Yildiz group, unpublished) [298–300]. Taken together, these results demonstrate that N protein can undergo bona fide phase separation, and that N protein condensates can form in cells. Nevertheless, the complexity introduced by multidimensional linkage effects in vivo could substantially influence the phase behavior and composition of condensates observed in the cell [346, 350, 373]. Of note, the regime we have identified in which phase separation occurs (Fig. 5.5) is remarkably relatively narrow, a prerequisite for the assembly of virion particles containing a single viral genome.

Does phase separation play a physiological role in SARS-CoV-2 biology? Phase separation has been invoked or suggested in many different viral contexts to date [374–378]. In SARS-CoV-2, one possible model suggests phase separation may drive recruitment of components to viral replication sites, although how this dovetails with the fact that replication occurs in double-membrane bound vesicles (DMVs) remains to be explored [300, 379]. An alternative

(and non-mutually exclusive) model is one in which phase separation catalyzes nucleocapsid polymerization, as has been proposed in elegant work on measles virus [343]. Here, the process of phase separation is decoupled from genome packaging, where gRNA condensation occurs through association with a helical nucleocapsid. If applied to SARS-CoV-2, such a model would suggest that (1) initially N protein and RNA phase separate in the cytosol, (2) some discrete pre-capsid state forms within condensates and, (3) upon maturation, the pre-capsid is released from the condensate and undergoes subsequent virion assembly by interacting with the membrane-bound M, E, and S structural proteins at the ER-Golgi intermediate compartment (ERGIC). While this model is attractive it places a number of constraints on the physical properties of this pre-capsid, not least that the ability to escape the "parent" condensate dictates that the assembled pre-capsid must interact less strongly with the condensate components than in the unassembled state. This requirement introduces some thermodynamic complexities: how is a pre-capsid state driven to assemble if it is necessarily less stable than the unassembled pre-capsid, and how is incomplete or abortive pre-capsid formation avoided if – as assembly occurs – the pre-capsid becomes progressively less stable?

A phase separation and assembly model raises additional questions, such as the origins of specificity for recruitment of viral proteins and viral RNA, the kinetics of pre-capsid-assembly within a large condensate, and preferential packaging of gRNA over sub-genomic RNA. None of these questions are unanswerable, nor do they invalidate this model, but they should be addressed if the physiological relevance of large cytoplasmic condensates is to be further explored in the context of virion assembly.

Our preferred interpretation is that N protein has evolved to drive genome compaction for packaging (Fig. 5.7). In this model, a single-genome condensate forms through N protein gRNA interaction, driven by a small number of high-affinity sites. This (meta)-stable single-genome condensate undergoes subsequent maturation, leading to virion assembly. In this model, condensate-associated N proteins are in exchange with a bulk pool of soluble N protein, such that the interactions that drive compaction are heterogeneous and dynamic. Our model

provides a physical mechanism in good empirical agreement with data for N protein oligomerization and assembly [380–382]. Furthermore, the resulting condensate is then in effect a multivalent binder for M protein, which interacts with N directly, and may drive membrane curvature and budding in a manner similar to that proposed by Bergeron-Sandoval and Michnick (though with a different directionality of the force) and in line with recent observations from cryo electron tomography (cryoET) [379, 383–385].

An open question pertains to specificity of packaging gRNA while excluding other RNAs. One possibility is for two high-affinity N-protein binding sites to flank the 5' and 3' ends of the genome, whereby only RNA molecules with both sites are competent for compaction A recent map of N protein binding to gRNA has revealed high-affinity binding regions at the 5' and 3' ends of the gRNA, in good agreement with this qualitative prediction [298]. Alternatively only gRNA condensates may possess the requisite valency to drive virion budding through interaction with M at the cytoplasmic side of the ERGIC, offering a physical selection mechanism for budding.

Genome compaction through dynamic multivalent interactions would be especially relevant for coronaviruses, which have extremely large single-stranded RNA genomes. This is evolutionarily appealing, in that as the genome grows larger, compaction becomes increasingly efficient, as the effective valence of the genome is increased [348, 349]. The ability of multivalent disordered proteins to drive RNA compaction has been observed previously in various contexts [292, 386]. Furthermore, genome compaction by RNA binding protein has been proposed and observed in other viruses [382, 387, 388], and the SARS coronavirus N protein has previously been shown to act as an RNA chaperone, an expected consequence of compaction to a dynamic single-RNA condensate that accommodates multiple N proteins with a single RNA [292, 389]. Furthermore, previous work exploring the ultrastructure of phase separated condensates of G3BP1 and RNA through simulations and cryoET revealed a beads-on-a-string type architecture, mirroring recent results for obtained from cryoET of SARS-CoV-2 virions [339, 379].

N protein has been shown to interact directly with a number of proteins studied in the context of biological phase separation which may influence assembly in vivo [283, 298, 338, 345, 390]. In particular, G3BP1-an essential stress-granule protein that undergoes phase separation-was recently shown to co-localize with overexpressed N protein [300, 339, 345, 391]. G3BP1 interaction may be part of the innate immune response, leading to stress-granule formation, or alternatively N protein may attenuates the stress response by sequestering G3BP1, depleting the cytosolic pool, and preventing stress granule formation, as has been shown for HIV-1 [378].

Our model is also in good empirical agreement with recent observations made for other viruses [392]. Taken together, we speculate that viral packaging may -in general- involve an initial genome compaction through multivalent protein:RNA and protein:protein interactions, followed by a liquid-to-solid transition in cases where well-defined crystalline capsid structures emerge. Liquid-to-solid transitions are well established in the context of neurodegeneration with respect to disease progression [393–395]. Here we suggest nature is leveraging those same principles as an evolved mechanism for monodisperse particle assembly.

Regardless of if phase separated condensates form inside cells, all available evidence suggests phase separation is reporting on a physiologically important interaction that underlies genome compaction (Fig. 5.6L). With this in mind, from a biotechnology standpoint, phase separation may be a convenient readout for in vitro assays to interrogate protein:RNA interaction. Regardless of which model is correct, N protein:RNA interaction is key for viral replication. As such, phase separation provides a macroscopic reporter on a nanoscopic phenomenon, in line with previous work [338, 348, 396, 397]. In this sense, we believe the therapeutic implications of understanding and modulating phase separation here (and elsewhere in biology) are conveniently decoupled from the physiological relevance of actual, large phase separated “liquid droplets”, but instead offer a window into the underlying physical interactions that lead to condensate formation.

5.4.4 The physics of single polymer condensates

Depending on the molecular details, single-polymer condensates may be kinetically stable (but thermodynamically unstable, as in our model simulations) or thermodynamically stable. Delineation between these two scenarios will depend on the nature, strength, valency and anisotropy of the interactions. It is worth noting that from the perspective of functional biology, kinetic stability may be essentially indistinguishable from thermodynamic stability, depending on the lifetime of a metastable species.

It is also important to emphasize that at higher concentrations of N protein and/or after a sufficiently long time period we expect robust phase separation with viral RNA, regardless of the presence of a symmetry-breaking site. Symmetry breaking is achieved when the apparent local concentration of N protein (from the "perspective" of gRNA) is substantially higher than the actual global concentration. As effective local and global concentrations approach one another, the entropic cost of intra-molecular interaction is outweighed by the availability of inter-molecular partners. On a practical note, if the readout in question is the presence/absence of liquid droplets, a high-affinity site may be observed as a shift in the saturation concentration which, confusingly, could either suppress or enhance phase separation. Further, if single-genome condensates are kinetically stable and driven through electrostatic interactions, we would expect a complex temperature dependence, in which larger droplets are observed at higher temperature (up to some threshold). Recent work is showing a strong temperature-dependence of phase separation is consistent with these predictions [298].

Finally, we note no reason to assume single-RNA condensates should be exclusively the purview of viruses. RNAs in eukaryotic cells may also be processed in these types of assemblies, as opposed to in large multi-RNA RNPs. The role of RNA:RNA interactions both here and in other systems is also of particular interest and not an aspect explored in our current work, but we anticipate may play a key role in the relevant biology.

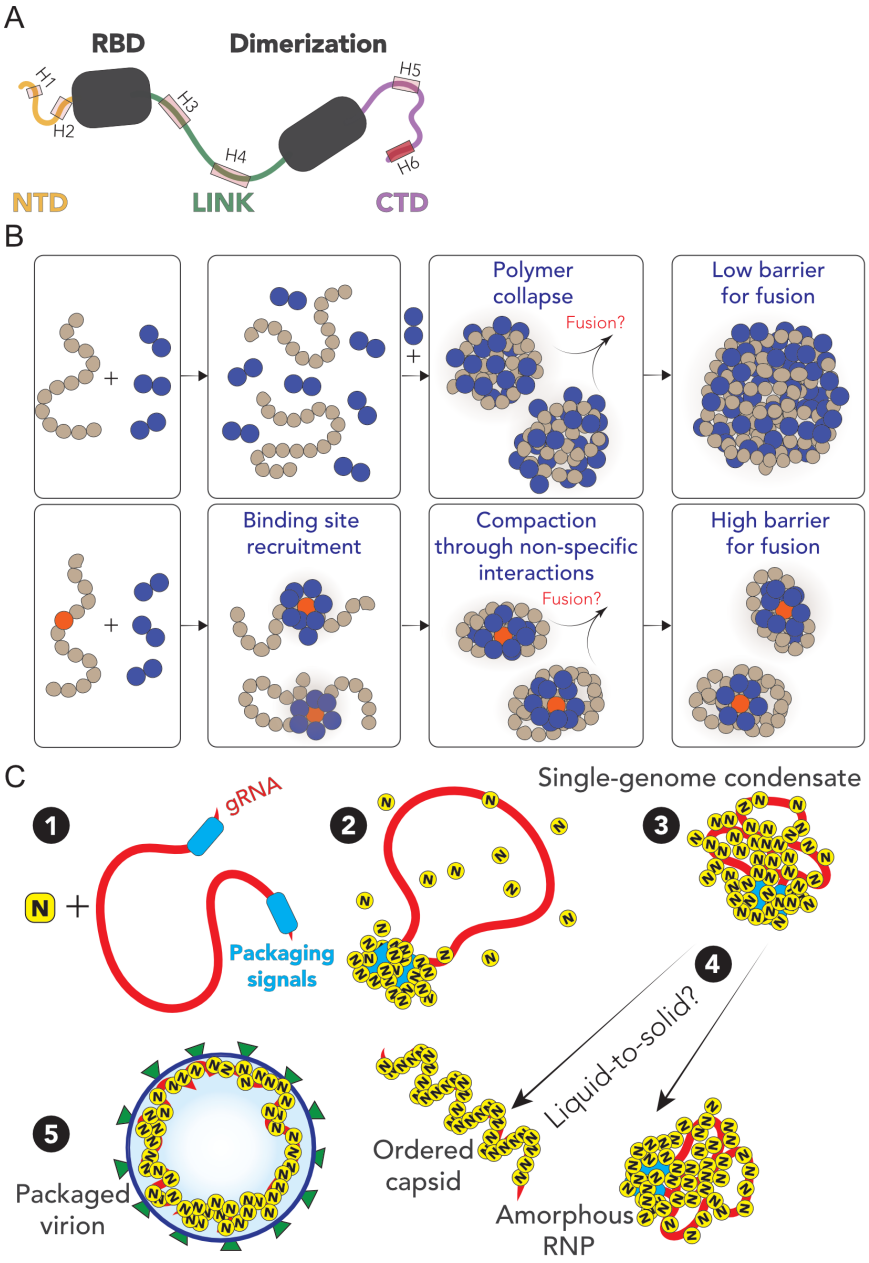


Figure 5.7: Summary and proposed model for N protein behavior. **A.** Summary of results from single-molecule spectroscopy experiments and all-atom simulations. All three predicted IDRs are disordered, highly flexible, and house a number of putative helical binding regions which overlap with subregions identified previously to drive N protein function. **B.** Overview of general symmetry breaking model. For homopolymers, local collapse leads to single-polymer condensates with a small barrier to fusion, rapidly assembling into large multi-polymer condensates. When one (or a small number of) high-affinity sites are present, local clustering of binders at a lower concentration organize the polymer such that single-polymer condensates are kinetically stable. **C.** Proposed model for SARS-CoV-2 genome packaging. **(1)** Simplified model of SARS-CoV-2 genome with a pair of packaging region at the 5' and 3' end of the genome **(2)** N protein preferentially binds to packaging signal regions in the genome, leading to a local cluster of N protein at the packaging signal RNA. **(3)** The high local concentration of N protein drives condensation of distal regions of the genome, forming a stable single-genome condensate. **(4)** Single-genome condensates may undergo subsequent maturation through a liquid-to-solid (crystallization) transition to form an ordered crystalline capsid, or solidify into an amorphous ribonuclear particle (RNP), or some combination of the two. While in some viruses an ordered capsid clearly forms, we favour a model in which the SARS-CoV-2 capsid is an amorphous RNP. Compact single-genome condensates ultimately interact with E, S and M proteins at the membrane, whose concerted action leads to envelope formation around the viral RNA and final virion packaging.

5.5 Methods

5.5.1 All atom simulations

All-atom Monte Carlo simulations were performed with the ABSINTH implicit solvent model and CAMPARI simulation engine (http://campari.sourceforge.net/) [398,399] with the solution ion parameters of Mao et al. [400]. Simulations were performed using movesets and Hamiltonian parameters as reported previously [338, 401]. All simulations were performed in sufficiently large box sizes to prevent finite size effects (where box size varies from system to system). For simulations with IDRs in isolation all degrees of freedom available in CAMPARI are sampled. For simulations with folded domains with IDRs, the backbone dihedral angles in folded domains are not sampled, such that folded domains remain structurally fixed (although sidechains are fully sampled). The IDR has backbone and sidechain degrees of freedom sampled.

All-atom molecular dynamics simulations were performed using GROMACS, using the FAST algorithm in conjunction with the Folding@home platform [35, 37, 222]. Post-simulation analysis was performed with Enspara [39]. For additional simulation details see the supplementary information.

5.5.2 Coarse-grained Polymer Simulations

Coarse-grained Monte Carlo simulations were performed using the PIMMS simulation engine [402]. All simulations were performed in a 70 x 70 x 70 lattice-site box. The results averaged over the final 20% of the simulation to give average values at equivalent states. The "polymer" is represented as a 61-residue polymer with either a central high-affinity binding site or not. The binder is a 2-bead species. Every simulation was run for 20 x 109 Monte Carlo steps, with four independent replicas. Bead interaction strengths were defined as shown in Fig. 5.6A. For

additional simulation details see the supplementary information.

5.5.3 Protein Expression, purification, and labeling.

SARS-CoV-2 Nucleocapsid protein (NCBI Reference Sequence: YP_009724397.2) including an N term extension containing His9-HRV 3C protease site was cloned into the BamHI EcoRI sites in the MCS of pGEX-6P-1 vector (GE Healthcare). Site-directed mutagenesis was performed on the His9-SARS-CoV-2 Nucleocapsid pGEX vector to create M1C R68C, Y172C T245C, and F363C A419C variant N protein constructs and sequences were verified using Sanger sequencing. All variants were expressed recombinantly in BL21 Codon-plus pRIL cells (Agilent) or Gold BL21(DE3) cells (Agilent) and purified using a FF HisTrap column. The GST-His9-N tag was then cleaved using HRV 3C protease and further purified to remove the cleaved tag. Finally, purified N protein variants were analyzed using SDS-PAGE and verified by electrospray ionization mass spectrometry (LC-MS). Activity of the protein was assessed by testing whether the protein is able to bind and condense nucleic acids (see phase-separation experiments) as well as to form dimers (see oligomerization in SI).

All Nucleocapsid variants were labeled with Alexa Fluor 488 maleimide and Alexa Fluor 594 maleimide (Molecular Probes) under denaturing conditions following a two-step sequential labeling procedure (see SI).

5.5.4 Single-molecule fluorescence spectroscopy.

Single-molecule fluorescence measurements were performed with a Picoquant MT200 instrument (Picoquant, Germany). FRET experiments were performed by exciting the donor dye with a laser power of 100 μW (measured at the back aperture of the objective). For pulsed interleaved excitation of donor and acceptor, the power used for exciting the acceptor dye was adjusted to match the acceptor emission intensity to that of the donor (between 50 and 70 mW).

Single-molecule FRET efficiency histograms were acquired from samples with protein concentrations between 50 pM and 100 pM and the population with stoichiometry corresponding to 1:1 donor:acceptor labeling was selected. Trigger times for excitation pulses (repetition rate 20 MHz) and photon detection events were stored with 16 ps resolution. For FRET-FCS, samples of double-labeled protein with a concentration of 100 pM were excited by either the diode laser or the supercontinuum laser at the powers indicated above.

All samples were prepared in 50 mM Tris pH 7.32, 143 mM β-mercaptoethanol (for photoprotection), 0.001% Tween 20 (for limiting surface adhesion) and GdmCl at the reported concentrations. All measurements were performed in uncoated polymer coverslip cuvettes (Ibidi, Wisconsin, USA), which significantly decrease the fraction of protein adhering to the surface (compared to normal glass cuvettes) under native conditions. For comparison, experiments have been performed also in glass cuvette coated with PEG, which provided analogous results to the polymeric cuvette. Each sample was measured for at least 30 min at room temperature (295 ± 0.5 K) (see appendix E).

5.6 Acknowledgements

We thank Amy Gladfelter, Christiane Iserman, Christine Roden, Ahmet Yildiz, Amanda Jack, Luke Ferro, Steve Michnick, Pascale Legault, and Jim Omichinski for sharing data and extensive discussion. We also thank Rohit Pappu for placing our groups in contact with one another.

We thank the labs of John Cooper, Carl Frieden, and Silvia Jansen for providing some of the reagents we have used in this work. We thank Ben Schuler and Daniel Nettels for developing, maintaining, and sharing with us the software package used to analyze the single-molecule data.

J.C. and J.J.A are supported by NIGMS R25 IMSD Training Grant GM103757. We are grateful to the citizen-scientists of Folding@home for donating their computing resources. G.R.B

holds an NSF CAREER Award MCB-1552471, NIH R01GM12400701, a Career Award at the Scientific Interface from the Burroughs Wellcome Fund, and a Packard Fellowship for Science and Engineering from The David and Lucile Packard Foundation. []

A.S.H. is a scientific consultant with Dewpoint Therapeutics.

Chapter 6

Citizen Scientists Create an Exascale Computer to Combat COVID-19

This chapter is adapted from the following publication:

Zimmerman, M.I., Porter, J.R., Ward, M.D., Singh, S., Vithani, N., Meller, A., Mallimadugula, U.L., Kuhn, C.E., Borowski, J.H., Wiewiora, R.P., Hurley, M.F.D., Coffland, J.E., Voelz, V.A., Chodera, J.D., Bowman, G.R., Available on BiorXiv at https://doi.org/10.1101/2020.06.27.175430 [403]

My work in setting up almost all SARS-CoV-2 system for simulation and managing the Folding@home network led figures 6.1 and 6.4 and the data described in table 6.1

6.1 Abstract

The SARS-CoV-2/COVID-19 pandemic continues to threaten global health and socioeconomic stability. Experiments have revealed snapshots of many of the viral components but remain blind to moving parts of these molecular machines. To capture these essential processes, over

a million citizen scientists have banded together through the Folding@home distributed computing project to create the world's first Exascale computer and simulate protein dynamics. An unprecedented 0.1 seconds of simulation of the viral proteome reveal how the spike complex uses conformational masking to evade an immune response, conformational changes implicated in the function of other viral proteins, and 'cryptic' pockets that are absent in experimental snapshots. These structures and mechanistic insights present new targets for the design of therapeutics.

6.2 Introduction

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is a novel coronavirus that poses an imminent threat to global human health and socioeconomic stability [404, 405]. With estimates of the basic reproduction number at 3-4 and a case fatality rate for coronavirus disease 2019 (COVID-19) ranging from 0.1-12% (high temporal variation), SARS-CoV-2/COVID-19 has the potential to spread quickly and endanger the global population [405–409]. As of June 23rd, 2020, there have been over 9.1 million confirmed cases and over 472,000 fatalities, globally. Quarantines and social distancing are effective at slowing the rate of infection; however, they cause significant social and economic disruption. Taken together, it is crucial that we find immediate therapeutic interventions.

A structural understanding of the SARS-CoV-2 proteins could accelerate the discovery of new therapeutics by enabling the use of rational design [410]. Towards this end, the structural biology community has made heroic efforts to rapidly build models of SARS-CoV-2 proteins and the complexes they form. However, it is well established that a protein's function is dictated by the full range of conformations it can access; many of which remain hidden to experimental methods. Mapping these conformations for proteins in SARS-CoV-2 will provide a clearer picture of how they accomplish their functions, such as infecting cells, evading the immune system, and replicating. Such maps may also present new therapeutic opportunities, such as

'cryptic' pockets that are absent in experimental snapshots but provide novel targets for drug discovery.

Molecular dynamics simulations have the ability to capture the full ensemble of structures a protein adopts but require significant computational resources. Such simulations capture an all-atom representation of the range of motions a protein undergoes. Modern datasets often consist of a few μs of simulation for a single protein, with a few noteworthy examples reaching ms timescales. However, many important processes occur on slower timescales. Moreover, simulating every protein that is relevant to SARS-CoV-2 for biologically relevant timescales would require compute resources on an unprecedented scale.

To overcome this challenge, more than a million citizen scientists from around the world have donated their computer resources to simulate SARS-CoV-2 proteins. This massive collaboration was enabled by the Folding@home distributed computing platform, which has crossed the Exascale computing barrier and is now the world's largest supercomputer. Using this resource, we constructed quantitative maps of the structural ensembles of over two dozen proteins and complexes that pertain to SARS-CoV-2. Together, we have run an unprecedented 0.1 s of simulation. Our data uncover the mechanisms of conformational changes that are essential for SARS-CoV-2's replication cycle and reveal a multitude of new therapeutic opportunities. The data are supported by a variety of experimental observations and have been made publicly available (https://covid.molssi.org/) in accordance with open science principles to accelerate the discovery of new therapeutics.

6.3 Results and discussion

6.3.1 To the Exascale and beyond!

Folding@home is a community of citizen scientists, researchers, and tech organizations that apply their collective computational and intellectual resources to understand the role of proteins' dynamics in their function and dysfunction, and to aid in the design of new proteins and therapeutics. The project was founded in the year 2000 with the intent of understanding how proteins fold. At the time, simulating the folding of even small proteins could easily take thousands of years on a single computer. To overcome this challenge, the scientific team created a way to break these intractable problems into small pieces that could be performed completely independently of one another. They then created the Folding@home project to enable anyone with a computer and an internet connection to volunteer to run these small chunks of simulation, called work units.

Over the years, Folding@home has generalized to address many aspects of protein dynamics, and the algorithms have developed significantly. The project has provided insight into diverse topics, ranging from signaling mechanisms [48, 411, 412] to the connection between phenotype and genotype [9, 10, 59]. Translational applications have included new means to combat antimicrobial resistance, Ebola virus, and SFTS virus [49, 54, 413, 414].

In response to the COVID-19 pandemic, Folding@home quickly pivoted to focus on SARS-CoV-2 and the host factors it interacts with. Many people found the opportunity to take action at a time when they were otherwise feeling helpless alluring. In less than three months, the project grew from 30,000 active devices to over a million devices around the globe (Fig. 6.1A and 6.1B).

Estimating the aggregate compute power of Folding@home is non-trivial due to factors like hardware heterogeneity, measures to maintain volunteers' anonymity, and the fact that vol-

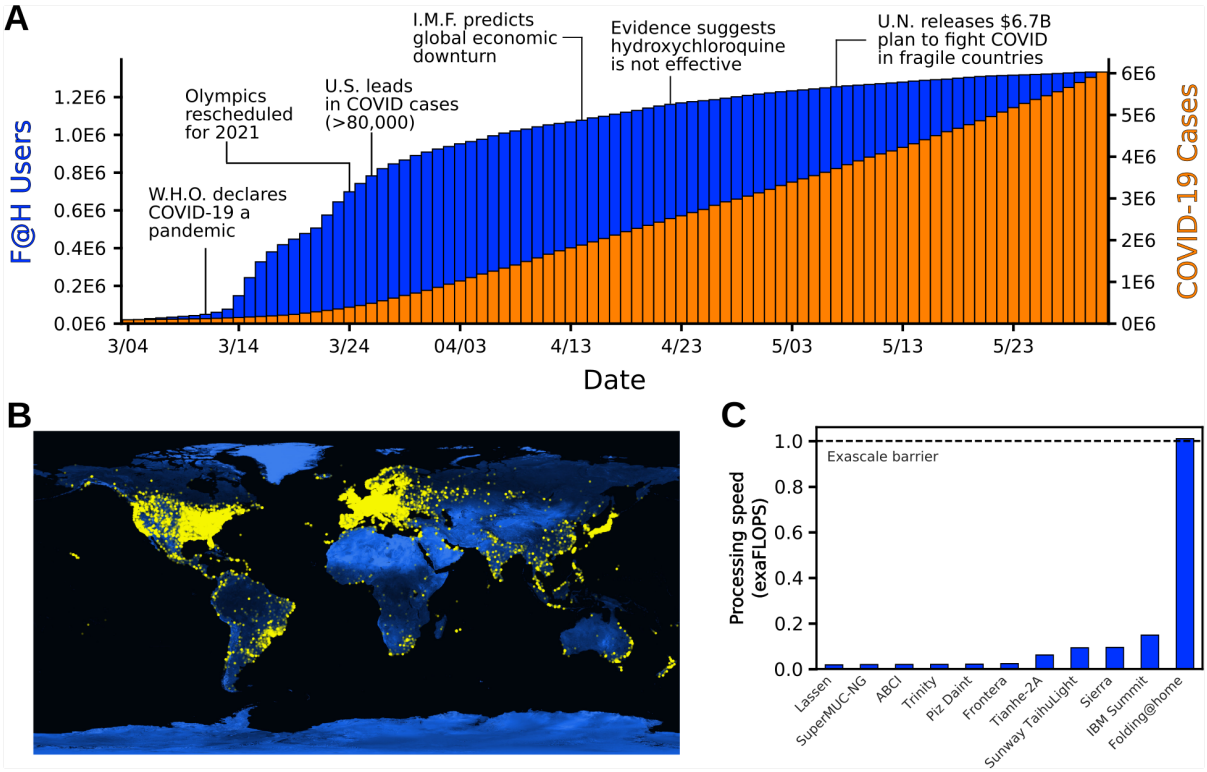


Figure 6.1: Summary of Folding@home's computational power. **A.** The growth and usage of Folding@home in response to COVID-19. Users are colored blue and COVID-19 cases are orange. **B.** Location of folding at home users. Each yellow dot represents a unique IP address contributing to Folding@home. **C.** The processing speed of Folding@home and the next 10 fastest supercomputers, in exaFLOPS.

unteers can turn their machines on and off at-will. Furthermore, volunteers' machines only communicate with the Folding@home servers at the beginning and end of a work unit, each of which can take anywhere from tens of minutes to a few days depending on the volunteer's hardware and the protein to simulate. Therefore, we chose to estimate the performance by counting the number of GPUs and CPUs that participated in Folding@home during a three-day window and making a conservative assumption about the computational performance of each device (see Methods for details).

Given the above, we estimate the peak performance of Folding@home hit 1.01 exaFLOPS. This performance was achieved at a point when 280,000 GPUs and 4.8 million CPU cores were performing simulations. For reference, that performance is 5-fold greater than the peak performance of the world's fastest traditional supercomputer, called Summit (Fig. 6.1C). It

is also more than the top 100 supercomputers combined. Prior to Folding@home, the first exascale supercomputer was not scheduled to come online until the end of 2021.

6.3.2 Unmasking the spike complex

The spike complex (S) is a prominent vaccine target that is known to undergo substantial conformational changes as part of its function [415–417]. Structurally, S is composed of three interlocking proteins, with each chain having a cleavage site separating an S1 and S2 fragment. S resides on the virion surface, where it waits to engage with an angiotensin-converting enzyme 2 (ACE2) receptor on a host cell to trigger infection [418, 419]. The fact that S is exposed on the virion surface makes it an appealing vaccine target. However, it has a number of effective defense strategies. First, S is decorated extensively with glycans that aid in immune evasion by shielding potential antigens [420]. S also uses a conformational masking strategy, wherein it predominantly adopts a closed conformation that buries the receptor-binding domains (RBDs) to evade immune surveillance mechanisms. To engage with ACE2, S undergoes rare transitions to an open state that exposes the conserved binding interface of the RBDs. Characterization of the full range of this motion is important for understanding pathogenesis and could provide insights into novel therapeutic options.

To capture S opening, we employed our goal-oriented adaptive sampling algorithm, FAST, in conjunction with Folding@home. The FAST method iterates between running a batch of simulations, building an MSM, ranking the MSM states based on how likely starting a new simulation from that state is to yield useful data, and starting a new batch of simulations from the top ranked states [35, 421]. The ranking function is designed to balance between favoring structures with a desired geometric feature (in this case opening of S) and broad exploration of conformational space. By balancing exploration-exploitation tradeoffs, FAST often captures conformational changes with orders of magnitude less simulation time than alternative methods. Broadly distributed structures from our FAST simulations were then used as starting

points for extensive Folding@home simulations, totaling 1 ms of data, enabling us to obtain a statistically sound final model.

Our SARS-CoV-2 S protein simulations capture opening of S and substantial conformational heterogeneity in the open state (Fig. 6.3). Capturing opening of S is an impressive technical feat given that previous large-scale simulations were unable to observe this essential event for the initiation of infection. Intriguingly, we find that opening occurs only for a single RBD at a time, akin to that observed in cryoEM structures [422]. Additionally, we find that the scale of this opening can be substantially larger than has been observed in experimental snapshots (Fig. 6.3). The dramatic opening we observe is consistent with the observation that antibodies can bind to regions of the RBD that are deeply buried and seemingly inaccessible in existing experimental snapshots [423].

To understand the potential role of conformational masking in determining the lethality and infectivity of different coronaviruses, we also simulated the opening of S proteins from two related viruses: SARS-CoV-1 and HCoV-NL63. These viruses were selected because they also bind the ACE2 receptor but are associated with varying mortality rates. SARS-CoV-1 caused an outbreak in 2003 with a high case fatality rate but has not become a pandemic [424]. NL63 was discovered the following year and continues to spread around the globe, although it is significantly less lethal than either SARS virus [425]. We hypothesized that these phenotypic differences may be explained by changes to the S conformational ensemble. Specifically, we propose mutations or other perturbations can increase the S-ACE2 affinity by increasing the probability that S adopts an open conformation or by increasing the affinity between an exposed RBD and ACE2.

As expected, the three S complexes have very different propensities to adopt an open state and bind ACE2. Structures from each ensemble were classified as competent to bind ACE2 if superimposing an ACE2-RBD structure on S did not result in any steric clashes between ACE2 and the rest of the S complex. We find that SARS-CoV-1 has the highest population

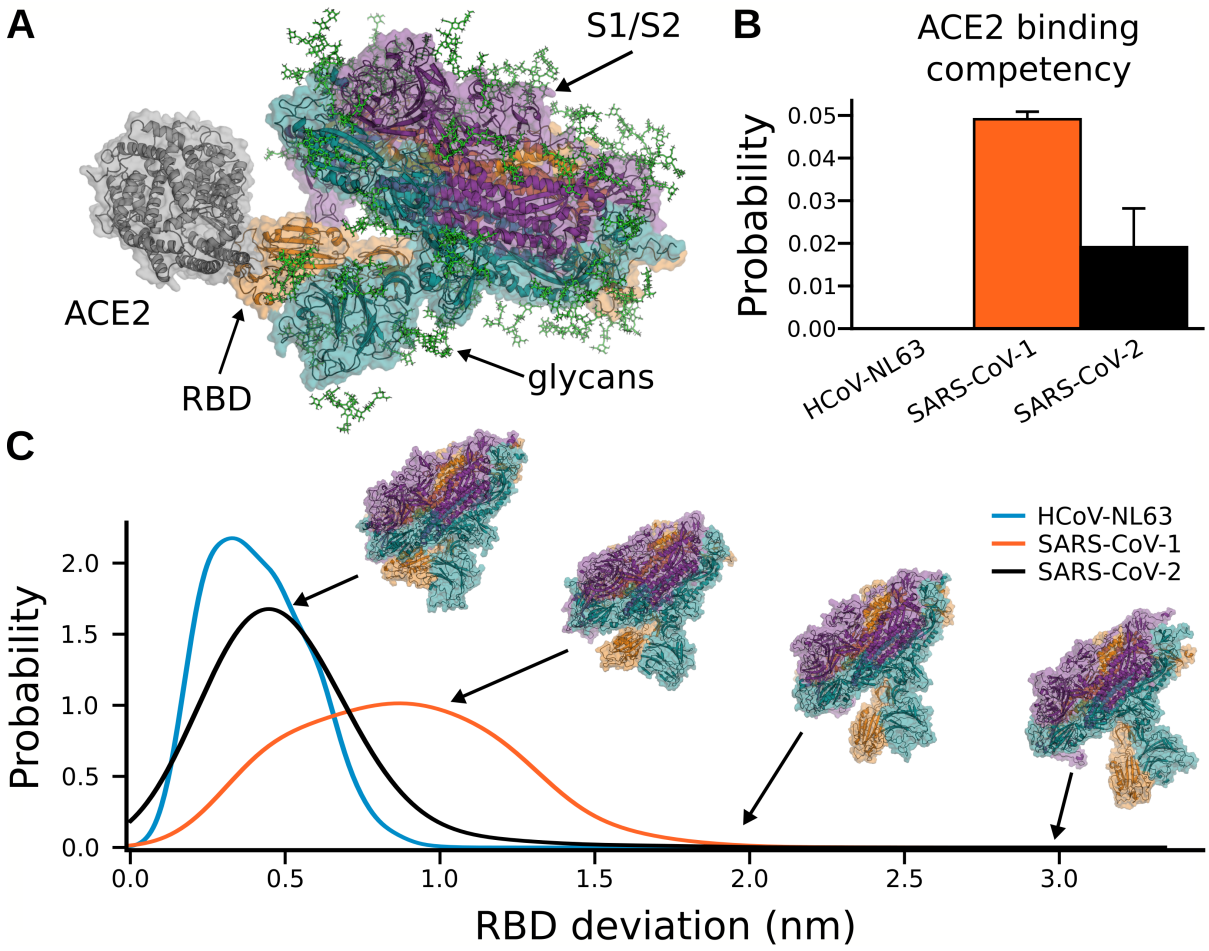


Figure 6.2: Structural characterization of conformational masking in different spike complexes. **A.** A representative structure of SARS-CoV-2 spike protein in an open conformation, as pulled from our molecular dynamic simulations. ACE2 (gray) is superimposed onto the structure to highlight binding compatibility. The three chains of Spike are illustrated with a cartoon and transparent surface representation (orange, teal, and purple), and glycans are shown as sticks (green). **B.** The probability that each sequence adopts an ACE2 binding competent pose. HCoV-NL63, SARS-CoV-1, and SARS-CoV-2 are shown as light-blue, orange, and black, respectively. **C.** The probability that the center of mass of an RBD deviates from its position in the closed state for HCoV-NL63, SARS-CoV-1, and SARS-CoV-2.

of conformations that can bind to ACE2 without steric clashes, followed by SARS-CoV-2, while opening of NL63 is sufficiently rare that we did not observe ACE2-binding competent conformations in our simulations (Fig. 6.2B). Interestingly, S proteins that are more likely to adopt structures that are competent to bind ACE2 are also more likely to adopt highly open structures (Fig. 6.2C).

We also observe a number of interesting correlations between conformational masking, lethal-

ity, and infectivity of different coronaviruses. First, more deadly coronaviruses have S proteins with less conformational masking. Second, there is an inverse correlation between S opening and the affinity of an isolated RBD for ACE2 (RBD-ACE2 affinities of 35 nM, 44 nM, and 185 nM for HCoV-NL63, SARS-CoV-2, and SARS-CoV-1, respectively) [426, 427].

These observations suggest a tradeoff wherein greater conformational masking enables immune evasion but requires a higher affinity between an exposed RBD and ACE2 to successfully infect a host cell. We propose that the NL63 S complex is probably best at evading immune detection but is not as infectious as the SARS viruses because strong conformational masking reduces the overall affinity for ACE2. In contrast, the SARS-CoV-1 S complex adopts open conformations more readily but is also more readily detected by immune surveillance mechanisms. Finally, SARS-CoV-2 balances conformational masking and the RBD-ACE2 affinity in a manner that allows it to evade an immune response while maintaining its ability to infect a host cell. Based on this model, we predict that mutations that increase the probability that the SARS-CoV-2 S complex adopts open conformations may be more lethal but spread less readily.

Our atomically detailed model of S can enable rapid structure-based vaccine antigen design through identification of regions minimally protected by conformational masking or the glycan shield [428]. To identify these potential epitopes, we calculated the probability that each residue in S could be exposed to therapeutics (e.g. not shielded by a glycan or buried by conformational masking), as shown in Fig. 6.3A. Visualizing these values on the protein reveals a few patches of protein surface that are exposed through the glycan shielding (Fig. 6.3B). However, another important factor when targeting an antigen is picking a region with a conserved sequence to yield broader and longer lasting efficacy. Not surprisingly, many of the exposed regions do not have a strongly conserved sequence. Promisingly, though, we do find a conserved area with a larger degree of solvent exposure (Fig. 6.3C). Another possibility for antigen design is to exploit the opening motion. A number of conserved residues of the RBD show an increase in exposure by $\tilde{3}0\%$ in ACE2 binding competent structures (Fig. 6.3C). Consistent with immunoassays, this region was recently found to be a cluster for neutralizing antibody

binding [423, 424].

6.3.3 Cryptic pockets and functional dynamics

Every protein in SARS-CoV-2 is a potential drug target. So, to understand their role in disease and help progress the design of antivirals, we unleashed the full power of Folding@home to simulate dozens of systems related to pathogenesis. While we are interested in all aspects of a proteins' functional dynamics, expanding on the number of antiviral targets is of immediate value. Towards this end, we seeded Folding@home simulations from our FAST-pockets adaptive sampling to aid in the discovery of cryptic pockets. We briefly discuss two illustrative examples, out of 36 datasets.

Nonstructural protein number 5 (NSP5, also named the main protease, Mpro, or 3CLpro) is critical for the lifecycle of coronaviruses and is a major target for the design of antivirals [429]. It is highly conserved between coronaviruses, owing to its necessary function of processing polyproteins. NSP5 is only active as a dimer, however it exists in a monomer-dimer equilibrium with estimates of its dissociation constant in the low μM range [430]. Small molecules targeting this protein to inhibit enzymatic activity, either by altering its active site or favoring the inactive monomer state, would be promising broad-spectrum antiviral candidates [431].

Our simulations reveal two novel cryptic pockets on NSP5 that expand our current therapeutic options. These are shown in figure 6.4A, which projects states from our MSM onto the solvent exposure of residues that make up the pockets. The first cryptic pocket is an expansion of NSP5's catalytic site. We find that the loop bridging domains II and III is highly dynamic and can fully disengage the protein. This motion may be necessary for catalysis and is similar to motions we have observed previously for the enzyme β-lactamase [49]. Owing to its location, a small molecule bound in this pocket is likely to prevent catalysis by obstructing polypeptide association with catalytic residues. The second pocket is a large opening between domains I/II and domain III. Located at the dimerization interface, this pocket offers the possibility to find

small molecule or peptide stabilizers that favor the inactive monomer state.

In addition to cryptic pockets, our data captures many potentially functionally relevant motions within the SARS-CoV-2 proteome. We illustrate this with the SARS-CoV-2 nucleoprotein. The nucleoprotein is a multifunctional protein responsible for major lifecycle events such as viral packaging, transcription, and physically linking RNA to the envelope [289, 301]. As such, we expect the protein to accomplish these goals through a highly dynamic and rich conformational ensemble, akin to context-dependent regulatory modules observed in Ebola virus nucleoprotein [2, 3]. Investigating the RNA-binding domain, we observe both cryptic pockets and an incredibly dynamic beta-hairpin, which hosts the RNA binding site, referred to as a "positive finger" (Fig. 6.4C-D). Our observed conformational heterogeneity of the positive finger is consistent with a structural ensemble determined using solution-state nuclear magnetic resonance spectroscopy [432]. Our simulations also capture numerous states of the putative RNA binding pose, where the positive finger curls up to form a cradle for RNA. These states can provide a structural basis for the design of small molecules that would compete with RNA binding, preventing viral assembly. Additionally, knowledge of these probabilities can provide further insight into the mechanisms and regulation of genome compaction/release.

The data we present in this paper represents the single largest collection of all-atom simulations. Table 6.1 is a comprehensive list of the systems we have simulated. Systems span various oligomerization states, include important complexes, and include representation from multiple coronaviruses. We also include human proteins that are targets for supportive therapies and preventative treatments. To accelerate the discovery of new therapeutics and promote open science, we are posting all of our data online (https://covid.molssi.org/).

Table 6.1: A list of protein systems we have simulated on Folding@home. Systems are organized by viral strain and include name, oligomerization state, starting structure, number of residues, number of atoms in the system, aggregate simulation time, and the number of cryptic pockets we have identified.
*Missing residues were modeled using Swiss model [1].
**Structural model was generated from a homologous sequence using Swiss model [1].
***Missing residues were modeled using CHARMM-GUI [2, 3]

System name	Oligomerization	Initial structure (homology)	Residues	Atoms in system	Aggregate simulation time (us)	Cryptic pockets discovered
SARS-CoV-2						
NSP3 (Macrodomain ÒXÓ)	Monomer	6W02	167	23907	10,531	TBD
NSP3 (Papain-like protease 2, PL2pro)	Monomer	3E9S**	306	97285	621	TBD
NSP5 (main protease, 3CLpro)	Monomer	6Y2E	306	64791	6,050	TBD
NSP5 (main protease, 3CLpro)	Dimer	6Y2E	612	77331	2,189	TBD
NSP7	Monomer	5F22**	79	20094	7,622	TBD
NSP8	Monomer	2AHM**	191	156282	1,259	TBD
NSP9	Dimer	6W4B*	226	49885	7,503	TBD
NSP10	Monomer	6W4H*	131	29560	3,050	TBD
NSP12 (polymerase)	Monomer	6NUR**	891	186622	1,114	TBD
NSP13 (helicase)	Monomer	6JYT**	596	129368	939	TBD
NSP14	Monomer	5C8S**	527	216380	359	TBD
NSP15	Monomer	6VWW	347	67345	3,618	TBD
NSP15	Hexamer	6VWW	2082	230339	1,461	TBD
NSP16	Monomer	6W4H*	298	45672	2,684	TBD
Nucleoprotein (RBD)	Monomer	6VYO	173	29125	8,742	TBD
Nucleoprotein Dimerization Domain	Monomer	6YUN*	118	34905	2,420	TBD
Nucleoprotein Dimerization Domain	Dimer	6YUN*	236	72733	1,251	TBD
Spike	Trimer	6VXX***	3363	442881	949	TBD
NSP7 / NSP8 / NSP12	Trimer complex	6NUR**	1184	215694	291	TBD
NSP10 / NSP14	Dimer complex	5C8S**	688	226672	469	TBD
NSP10 / NSP16	Dimer complex	6W4H*	429	63752	2,415	TBD
SARS-CoV-1						
NSP3 (Macrodomain ÒXÓ)	Monomer	2FAV	172	33117	507	TBD
NSP9	Dimer	1QZ8*	226	49599	6,736	TBD
NSP15	Monomer	2H85	345	67345	2,663	TBD
NSP15	Hexamer	2H85	2070	230339	829	TBD
Nucleoprotein RBD	Monomer	2OFZ	174	29125	4,124	TBD
Nucleoprotein Dimerization Domain	Monomer	2GIB	370	34905	874	TBD
Nucleoprotein Dimerization Domain	Dimer	2GIB	740	72733	427	TBD
Spike	Trimer	5X58***	3261	375851	784	TBD
NSP10 / NSP16	Dimer complex	6W4H**	425	69589	175	TBD
Human						
IL6	Monomer	1ALU	166	26855	1,263	TBD
IL6-R	Monomer	1N26	299	149764	132	TBD
ACE2	Monomer	6LZG	596	75787	244	TBD
MERS						
NSP13	Monomer	5WWP	596	121134	288	TBD
NSP10 / NSP16	Dimer Complex	6W4H**	424	69127	174	TBD
HCoV-NL63						
Spike	Trimer	5SZS***	3606	453348	618	TBD

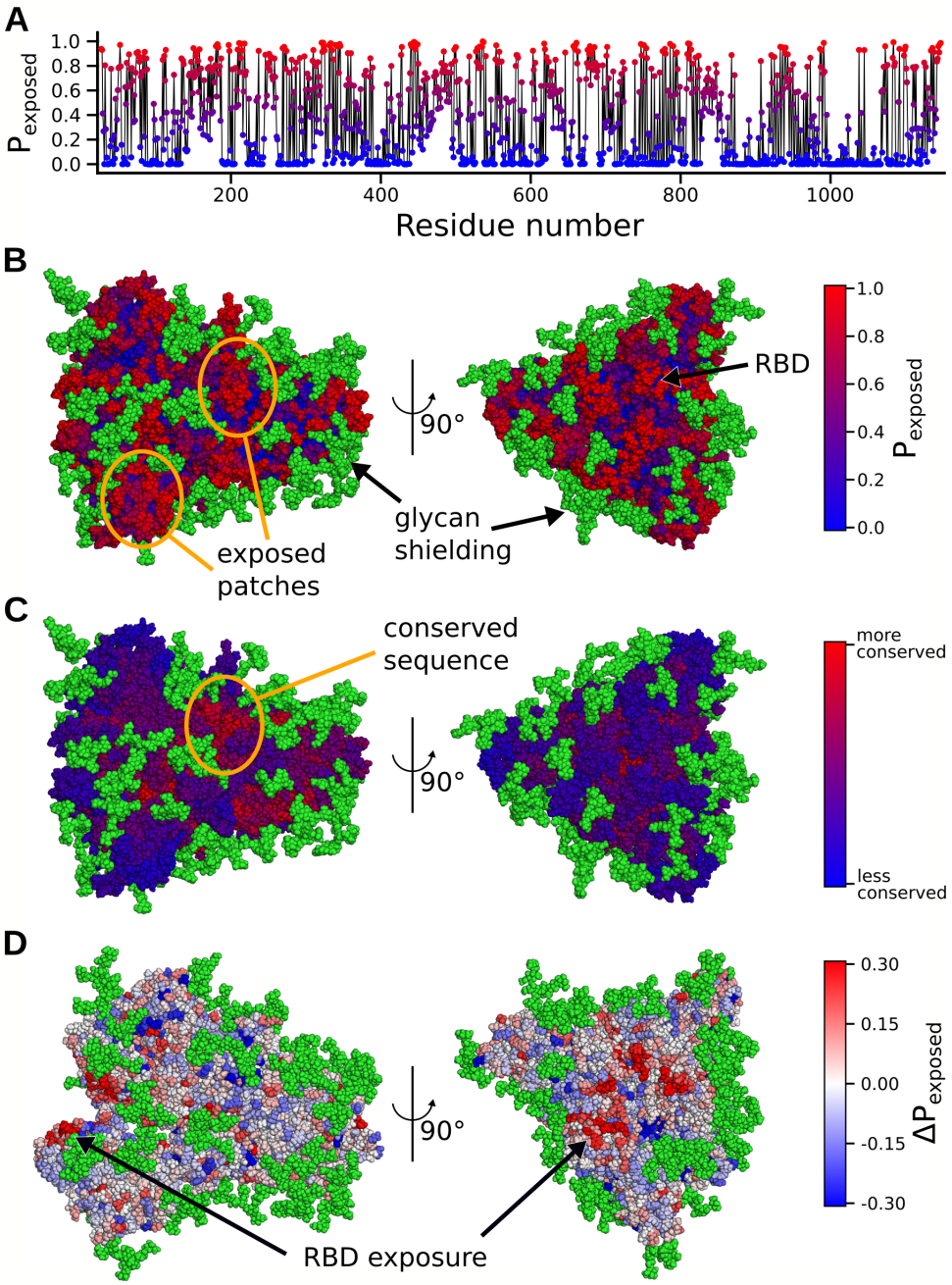


Figure 6.3: Effects of glycan shielding and conformational masking on the accessibility of different parts of the spike to potential therapeutics. **A.** The probability that a residue is exposed to potential therapeutics, as determined from our structural ensemble. **B.** Exposure probabilities colored on the surface of the spike protein. Exposed patches are circled orange. Red residues have a higher probability of being exposed, whereas blue residues have a lower probability of being exposed. **C.** Sequence conservation score colored onto the Spike protein. A conserved patch on the protein is circled in orange. Red residues have higher conservation, whereas blue residues have lower conservation. **D.** The difference in the probability that each residue is exposed between the ACE2-binding competent conformations and the entire ensemble. Red residues have a higher probability of being exposed upon opening, whereas blue residues have a lower probability of being exposed.

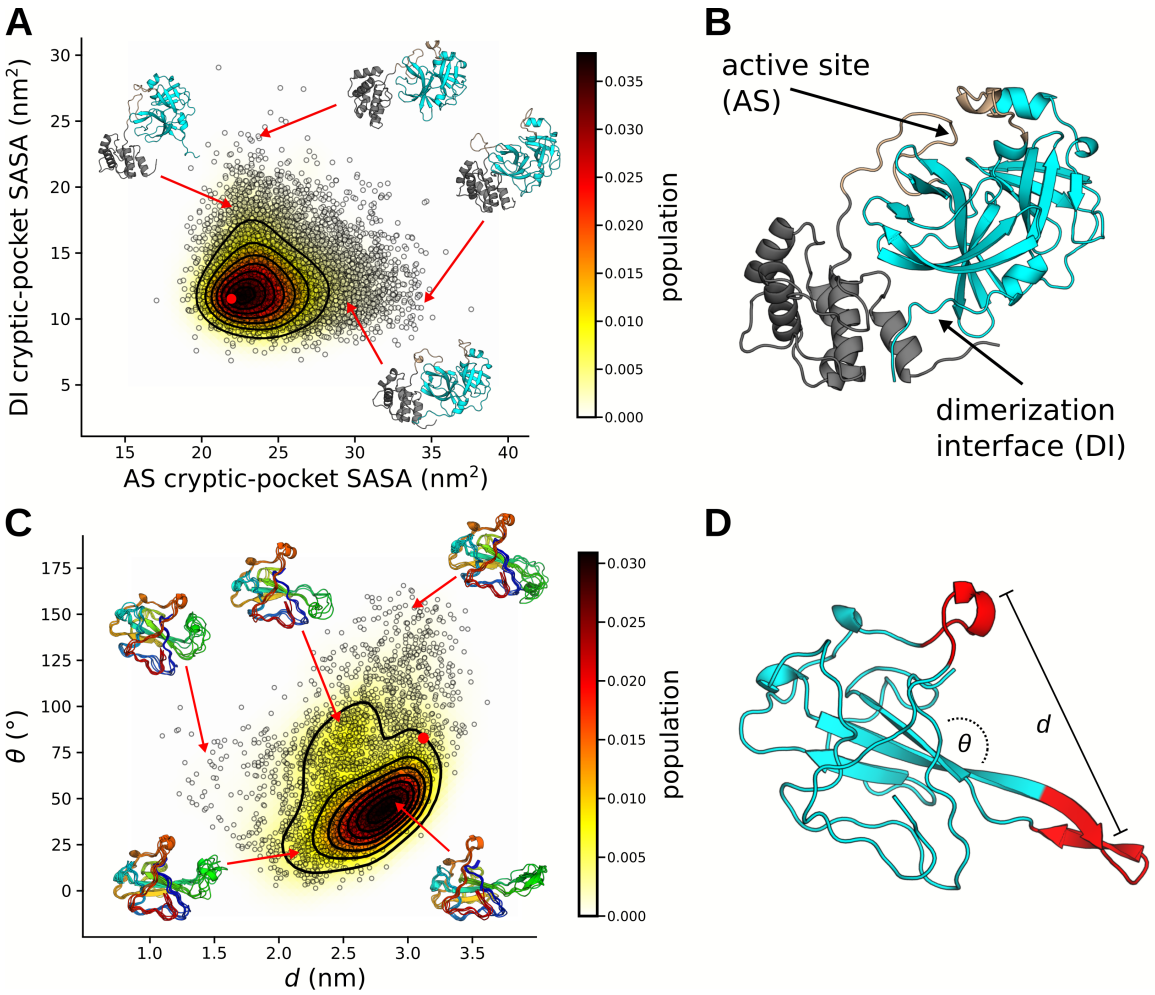


Figure 6.4: Examples of cryptic pockets and functionally-relevant dynamics. **A-B)** Conformational ensemble of the main protease (NSP5) projected onto the solvent accessible surface areas (SASAs) of residues surrounding either the active-site or a cryptic pocket. Cluster centers are represented as black circles with representative structures depicted with cartoon. The starting structure for simulations (6Y2E) is shown as a red dot. Domains I and II are colored cyan and domain III is colored gray. The loop of domain III, that covers the active-site residues and is seen to be highly dynamic, is colored tan. **C-D)** Conformational ensemble of Nucleoprotein projected onto the distance and angle between the positive finger and a nearby loop. Angles were calculated between vectors that point along each red segment in panel D and distances were calculated between their centers of mass. Cluster centers are represented as black circles and representative structures are depicted with cartoon. The starting structure for simulations (6VYO) is shown as a red dot.

6.4 Discussion

In this work, we have utilized the largest computer in the world to tackle a global threat. The pandemic caused by SARS-CoV-2 has necessitated a call-to-arms; a call that over a million citizen-scientists have answered, generating more than 0.1 seconds of simulation data. The unprecedented scale of these simulations has helped to characterize crucial stages of infection. We find that spike proteins have a strong trade-off between making ACE2 binding interfaces accessible to infiltrate cells and conformationally masking epitopes to subvert immune responses. SARS-CoV-2 represents a more optimal tradeoff than related coronaviruses, which may explain its success in spreading globally. Our simulations also provide an atomically detailed roadmap for targeting proteins for vaccines and antivirals. Furthermore, we are working on making a comprehensive repository of cryptic pockets hosted online to accelerate the development of novel therapeutics.

Beyond SARS-CoV-2, we expect this work to aid in a better understanding of the roles of proteins in the coronaviridae family. Coronaviruses have been around for millennia, yet many of their proteins are still poorly understood. Because climate change has made zoonotic transmission events more commonplace, it is imperative that we continue to perform basic research on these viruses to better protect us from future pandemics. For each protein system in Table 6.1, an extraordinary amount of sampling has led to the generation of a quantitative map of its conformational landscape. There is still much to learn about coronavirus function and these conformational ensembles contain a wealth of information to pull from.

While we have aggressively targeted research on SARS-CoV-2, Folding@home is a general platform for running molecular dynamics simulations at scale. Before the COVID-19 pandemic, Folding@home was already generating datasets that were orders of magnitude greater than from conventional means. With our explosive growth, our compute power has increased around 100-fold. Our work here highlights the incredible utility this compute power has to rapidly understand health, disease, and aid in drug design. Both in terms of scale and approach,

we are in a new frontier of using molecular dynamics simulations to understand biophysics; the complex task of simulating an organism's entire proteome could become commonplace. With the continued support of the citizen scientists that have made this work possible, we have the opportunity to make a profound impact on other global health crises such as cancer, neurodegenerative diseases, and antibiotic resistance.

6.5 Methods

6.5.1 System preparations

All simulations were prepared using Gromacs 2020 [222]. Initial structures were placed in a dodecahedron box that extends 1.0 nm beyond the protein in any dimension. Systems were then solvated and energy minimized with a steepest descents algorithm until the maximum force fell below 100 kJ/mol/nm using a step size of 0.01 nm and a cutoff distance of 1.2 nm for the neighbor list, Coulomb interactions, and van der Waals interactions. The AMBER03 force field was used for all systems except spike protein with glycans, which used CHARMM36 [146, 433]. All simulations were simulated with explicit TIP3P solvent [143].

Systems were then equilibrated for 1.0 ns, where all bonds were constrained with the LINCS algorithm and virtual sites were used to allow a 4 fs time step [226, 228]. Cutoffs of 1.1 nm were used for the neighbor list with 0.9 for Coulomb and van der Waals interactions. The Verlet cutoff scheme was used for the neighbor list. The stochastic velocity rescaling (v-rescale) thermostat was used to hold the temperature at 300 K [224].

6.5.2 Adaptive sampling simulations

The FAST algorithm was employed for each protein in Table 6.1 to enhance conformational sampling and quickly explore dominant motions. The procedure for FAST simulations is as follows: 1) run initial simulations, 2) build MSM, 3) rank states based on FAST ranking, 4) restart simulations from the top ranked states, 5) repeat steps 2-4 until ranking is optimized. For each system, MSMs were generated after each round of sampling using a k-centers clustering algorithm based on the RMSD between select atoms. Clustering continued until the maximum distance of a frame to a cluster center fell within a predefined cutoff. In addition to the FAST ranking, a similarity penalty was added to promote conformational diversity in starting structures, as has been described previously [278].

FAST-distance simulations of all Spike proteins were run at 310 K on the Microsoft Azure cloud computing platform. The FAST-distance ranking favored states with greater RBD openings using a set of distances between atoms. Each round of sampling was performed with 22 independent simulations that were 40 ns in length (0.88 μs aggregate sampling per round), where the number of rounds totaled 13 (11.44 μs), 22 (19.36 μs), and 17 (14.96 μs), for SARS-CoV-1, SARS-CoV-2, and HCoV-NL63, respectively.

For all other proteins, FAST-pocket simulations were run at 300 K for 6 rounds, with 10 simulations per round, where each simulation was 40 ns in length (2.4 μs aggregate simulation). The FAST-pocket ranking function favored restarting simulations from states with large pocket openings. Pocket volumes were calculated using the LIGSITE algorithm [434].

6.5.3 Folding@home simulations

For each adaptive sampling run, a conformationally diverse set of structures was selected to be run on Folding@home. Structures came from the final k-centers clustering of adaptive sampling, as is described above. Simulations were deployed using a simulation core based on

either GROMACS 5.0.4 or OpenMM 7.4.1 [34, 222].

6.5.4 Markov state models

A Markov state model is a network representation of a free energy landscape and is a key tool for making sense of molecular dynamics simulations [38]. All MSMs were built using our python package, enspara [39]. Each system was clustered with the combined FAST and Folding@home datasets. In the case of spike proteins, states were defined geometrically based on the RMSD between backbone Cα coordinates. States were generated as the top 3000 centers from a k-centers clustering algorithm. All other proteins were clustered based on the euclidean distance between the solvent accessible surface area of residues, as is described previously [49]. Systems generated either 2500, 5000, 7500, or 10000 cluster centers from a k-centers clustering algorithm. Select systems were refined with 1-10 k-medoid sweeps. Transition probability matrices were produced by counting transitions between states, adding a prior count of $1/n_states$, and row-normalizing, as is described previously [42]. Equilibrium populations were calculated as the eigenvector of the transition probability matrix with an eigenvalue of one.

6.5.5 Spike/ACE2 binding competency

To determine Spike protein binding competency to ACE2 the following structures of the RBD bound to ACE2 were used: 3D0G, 6M0J, and 3KBH, for SARS-CoV-1, SARS-CoV-2, and HCoV-NL63, respectively. The RBD of the bound complex was superimposed onto each RBD for structures in our MSM. Steric clashes were then determined between backbone atoms on the ACE2 molecule and the rest of the spike protein. If any of the structures had a superposition that resulted in no clashes, it was deemed binding competent.

6.5.6 Cryptic pockets and solvent accessible surface area

For ease of detecting cryptic pockets and other functional motions, we employed our exposon analysis method [49]. This method correlates the solvent exposure between residues to find concerted motions that tend to represent cryptic pocket openings. Solvent accessible surface area calculations were computed using the Shrake-Rupley algorithm as implemented in the python package MDTraj [148]. For all proteins and complexes, a solvent probe radius of 0.28 nm was used, which has been shown to produce a reasonable clustering and exposon map [49].

Spike protein solvent accessible surface areas for SARS-CoV-2 were computed with glycan chains modeled onto each cluster center. Multiple glycan rotamers were sampled for each state and accessible surface areas for each residue were weighted based on MSM equilibrium populations.

6.5.7 Sequence conservation

Sequence conservation of spike proteins was calculated using the Uniprot database [435]. Sequences between 30% - 90% were pulled and aligned with the Muscle algorithm [436]. The entropy at each position was calculated to quantify variability of amino acids. Conservation was defined as one minus the entropy [237].

6.6 Acknowledgements

We are extremely grateful to all the citizen scientists who contributed their compute power to make this work possible, and members of the Folding@home community who volunteered to help with everything from technical support to translating content into multiple languages. Thanks to Microsoft AI for Health for helping us use Azure to run adaptive sampling simulations, and to UKRI for providing compute resources to parallelize data analysis. Thanks

to Pure Storage for providing a FlashBlade system to store our large datasets, to Seagate and Micron for additional storage, and to MolSSI for helping organize public datasets. Thanks to Avast, AWS, Cisco, Linus Tech Tips, Microsoft Azure, Oracle, and VMware for helping us to scale-up Folding@home's server-side infrastructure to keep up with the tremendous growth we experienced in such a short time. Thanks to AMD, ARM and Neocortix, and Intel for helping to improve the performance of Folding@home on their hardware. Thanks to all of the above companies for spreading the word about Folding@home, and also to A16Z, Best Buy, CCP, CoreWeave, Daimler Truck AG, Dell, GitHub, HP, La Liga, Media Monks, Microcenter, NVIDIA, and Telefonica. Thanks to CERN and the particle physics community for providing guidance on data management strategies and to DataDog for server monitoring services. GRB and his lab were supported by funding from Avast, the Center for the Science and Engineering of Living Systems (CSELS), an NSF RAPID award, NSF CAREER Award MCB-1552471, NIH R01 GM124007, a Burroughs Wellcome Fund Career Award at the Scientific Interface, and a Packard Fellowship for Science and Engineering. JDC acknowledges support from NIH grant P30 CA008748 and NIH grant R01 GM121505. VAV and MFDH acknowledge support from NIH grant R01 GM123296.

6.7 Disclosures

JDC is a current member of the Scientific Advisory Board of OpenEye Scientific Software and a consultant to Foresite Laboratories. The Chodera laboratory receives or has received funding from multiple sources, including the National Institutes of Health, the National Science Foundation, the Parker Institute for Cancer Immunotherapy, Relay Therapeutics, Entasis Therapeutics, Silicon Therapeutics, EMD Serono (Merck KGaA), AstraZeneca, Vir Biotechnology, Bayer, XtalPi, the Molecular Sciences Software Institute, the Starr Cancer Consortium, the Open Force Field Consortium, Cycle for Survival, a Louis V. Gerstner Young Investigator Award, and the Sloan Kettering Institute. A complete funding history for the Chodera lab can

be found at http://choderalab.org/funding.

Chapter 7

Antagonism between substitutions in β-lactamase explains a path not taken in the evolution of bacterial drug resistance

This chapter is adapted from the following publication:

Brown, C.A., Hu, L., Sun, Z., Patel, M.P., Singh, S., Porter, J.R., Sankaran, B., Venkataram Prasad, B.V.V., Bowman, G.R., Palzkill, T., Antagonism between substitutions in β-lactamase explains a path not taken in the evolution of bacterial drug resistance, J.Biol., Chem., 2020, doi:10.1074/jbc.RA119.012489 [437]

My role in parameterizing and running molecular simulations of the crystal structures in apo and acyl-enzyme forms led to the data presented in figures 7.9 and 7.10, and appendix figures E.1, E.2, E.3.

7.1 Abstract

CTX-M β-lactamases are widespread in Gram-negative bacterial pathogens and provide resistance to the cephalosporin cefotaxime but not to the related antibiotic ceftazidime. Nevertheless, variants have emerged that confer resistance to ceftazidime. Two natural mutations, causing P167S and D240G substitutions in the CTX-M enzyme, result in 10-fold increased hydrolysis of ceftazidime. Although the combination of these mutations would be predicted to increase ceftazidime hydrolysis further, the P167S/D240G combination has not been observed in a naturally occurring CTX-M variant. Here, using recombinantly expressed enzymes, minimum inhibitory concentration measurements, steady-state enzyme kinetics, and X-ray crystallography, we show that the P167S/D240G double mutant enzyme exhibits decreased ceftazidime hydrolysis, lower thermostability, and decreased protein expression levels compared with each of the single mutants, indicating negative epistasis. X-ray structures of mutant enzymes with covalently trapped ceftazidime suggested that a change of an active-site Ω-loop to an open conformation accommodates ceftazidime leading to enhanced catalysis. 10-μs molecular dynamics simulations further correlated -loop opening with catalytic activity. We observed that the WT and P167S/D240G variant with acylated ceftazidime both favor a closed conformation not conducive for catalysis. In contrast, the single substitutions dramatically increased the probability of open conformations. We conclude that the antagonism is due to restricting the conformation of the -loop. These results reveal the importance of conformational heterogeneity of active-site loops in controlling catalytic activity and directing evolutionary trajectories.

7.2 Introduction

Enzymes have evolved to catalyze reactions critical to the functioning of the cell [438]. Evolution of enzyme function proceeds through the accumulation of amino acid substitutions that shape stability, solubility, and catalytic activity, among other properties. How substitutions

interact when combined plays a key role in the trajectory of mutations that accumulate during evolution [439, 440]. For example, amino acid substitutions can act additively on catalysis whereupon each substitution increases activity, and upon combination, the increase in activity in the double mutant is the product of the fold changes of the individual mutations [441]. Alternatively, combinations of substitutions are often nonadditive where the double mutant has a greater activity or less activity than expected based on the activity of the single mutants. Such nonadditive effects are termed epistasis and can strongly influence the mutational pathways that are possible in the evolution of enzyme function [442].

Enzymes act by binding substrates and stabilizing transition states of reactions [438]. Toward this end, conformational changes are often important, and flexible loops in the active site are a common feature involved in enzyme function [443–445]. Moreover, conformational dynamics have been proposed to play an important role in protein evolvability [446, 447]. By this view, conformational fluctuations can result in an enzyme adopting multiple structures, some of which have properties that allow interactions with alternate ligands. These conformations may be rare in the ensemble of WT structures, but mutations may shift the distribution toward alternate conformations that become dominant in an evolved enzyme, thereby allowing for altered substrate specificity or new enzyme functions to emerge on an enzyme scaffold [447,448].

Here, we address the role of epistasis and conformational diversity of active-site loops in the evolution of variants of the CTX-M β-lactamase with a broadened substrate specificity for β-lactam antibiotics. β-Lactams are the most frequently prescribed class of antibiotic worldwide, making up 65% of all use [449]. However, bacterial resistance to these drugs is a growing problem, and the most common mechanism of resistance is enzyme-mediated hydrolysis of the β-lactam ring [450]. This hydrolysis is catalyzed by various β-lactamases, which are divided into classes A–D based on primary amino acid sequence homology [450, 451].

Class A β-lactamases, such as CTX-M, are widespread in Gram-negative bacteria and share a similar mechanism of catalysis but can differ widely in substrate profile [452, 453]. These

enzymes are serine hydrolases that hydrolyze the amide bond in the β-lactam ring via sequential acylation and deacylation steps. The conserved catalytic Ser70 residue is activated by Lys73 and Glu166 for attack on the carbonyl carbon to form an acyl-enzyme intermediate. A catalytic water molecule is then activated by Glu166 for attack on the carbonyl of the covalent complex to deacylate the enzyme and release the product (Fig. 7.1) [453–455]. The reaction scheme and mechanism of serine β-lactamases is shown in Fig. 7.1.

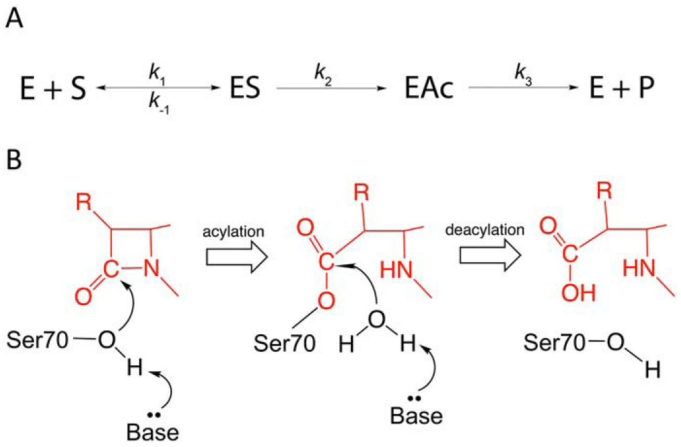


Figure 7.1: β-Lactamase mechanism. A, reaction scheme for $\beta-$lactamase where E represents $\beta-$lactamase, ES represents the enzyme-substrate complex, EAc represents the acyl-enzyme complex, and P represents product. k1 and k-1 are the rate constants for association and dissociation of the enzyme substrate complex, and k2 and k3 are the rate constants for acylation and deacylation, respectively. B, schematic illustration of $\beta-$lactamase mechanism. The catalytic Ser70 hydroxyl group is activated for nucleophilic attack on the carbonyl oxygen of the amide bond of the $\beta-$lactam by an active-site residue serving as a general base. This residue is viewed as either Lys73 or Glu166 acting through a water molecule. This leads to formation of the acyl enzyme intermediate, which is subsequently deacylated by a water that is activated by Glu166 acting as a base and resulting in free enzyme and the hydrolyzed product.

CTX-M β-lactamases are a family of class A extended-spectrum β-lactamases that are so named because they efficiently hydrolyze the oxyimino-cephalosporin cefotaxime [456] (Fig. 7.2). To date, more than 140 variants of the CTX-M enzymes have been identified [457]. CTX-M-14 β-lactamase has become a model system for studies of the structure and function of CTX-M enzymes [458–461].

CTX-M enzymes efficiently hydrolyze cefotaxime but not another commonly used oxyimino-cephalosporin, ceftazidime (Fig. 7.2A). Natural variants containing either the P167S or D240G

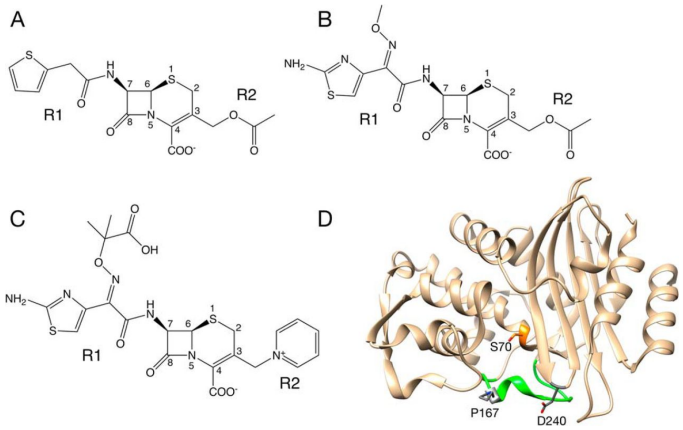


Figure 7.2: Structures of antibiotics and CTX-M-14 β-lactamase. A, cephalothin. R1 and R2 denote the groups that differ between cephalosporins. B, cefotaxime. C, ceftazidime. D, structure of CTX-M-14 β-lactamase (PDB code 1YLT). The Ω-loop is colored green. The catalytic Ser70 is colored orange. Pro167 and Asp240 are colored gray. Note that Pro167 is located on the Ω-loop.

substitutions have emerged, however, that more efficiently hydrolyze ceftazidime [457, 462–464]. These two substitutions, when present, individually increase k_{cat}/K_m for ceftazidime hydrolysis by 10-fold, resulting in increased ceftazidime resistance for bacteria containing the mutants [458,461]. Multiple natural variants in the CTX-M family possess one of these substitutions [457].

Pro167 resides in the Ω-loop that forms the bottom of the active site in class A β-lactamases including CTX-M enzymes [458,465] (Fig. 7.2D). It is adjacent to Glu166, which is conserved and serves as a general base to activate a water molecule for deacylation of β-lactam substrates [465]. The peptide bond preceding Pro167 is in a cis conformation in CTX-M enzymes, which strongly influences the conformation of the Ω-loop and the positioning of the Asn170 residue that hydrogen bonds to Glu166 and the deacylation water. We previously used the CTX-M-14 enzyme as a model system to examine the structural changes caused by the P167S substitution [461]. These studies revealed a large conformational change of the Ω-loop that results in a larger active-site cavity to accommodate ceftazidime. This conformational change required both the P167S substitution and the presence of acylated ceftazidime [461] In addition, the structures showed that the conformational change is associated with a shift in the peptide bond preceding residue 167 from cis to trans and that the P167S substitution was required for this

shift. Thus, the P167S substitution appears to cause increased ceftazidime hydrolysis through promoting a conformational change to relieve steric restraints on catalysis.

Chen et al. [458] previously determined the X-ray structure of the D240G mutant enzyme, and anisotropic B-factor analysis revealed increased flexibility of the B3 β-strand that forms one side of the CTX-M active site. The increased flexibility of the B3 β-stand was proposed to allow access for the bulky side chain of ceftazidime.

Despite the increase in ceftazidime hydrolysis and bacterial resistance resulting from each of the substitutions, there has yet to be a CTX-M enzyme identified in clinical isolates that harbors both the P167S and D240G mutations. Based on simple additivity, the combination of substitutions that each increase hydrolysis by 10-fold would be expected to increased hydrolysis 100-fold relative to the WT enzyme [441]. However, a P167S/D240G double mutant created by site-directed mutagenesis in a CTX-M-3 enzyme background exhibited a loss of ceftazidime resistance, indicating an antagonist effect and negative epistasis [466]. The mechanism of this antagonism, however, was not examined.

Here, we show that the P167S/D240G double mutant displays decreased ceftazidime hydrolysis compared with either of the single mutants, indicating antagonism. Further, X-ray structures of single and double mutants as apoenzymes and acylated with ceftazidime show alternate open and closed conformations of the Ω-loop that are associated with high and low activity. Finally, molecular dynamics simulations of the WT, P167S, D240G, and P167S/D240G enzymes acylated with ceftazidime indicate that the single substitutions dramatically increase the probability of open conformations of the Ω-loop, whereas the WT and P167S/D240G variant both favor a well-defined closed conformation not favorable for catalysis. Taken together, the results suggest that the P167S/D240G double mutant has not been observed in resistant clinical isolates because the combination results in decreased catalysis, decreased stability, and therefore decreased fitness in the presence of ceftazidime for bacteria containing this enzyme.

Table 7.1: MICs for E. coli containing CTX-M-14 wild type, mutants, and no $\beta-$lactamase control

	MIC (μg/ml)		
	Cephalothin	Cefotaxime	Ceftazidime
pTP123	12	0.0625	0.19
CTX-M-14 wt	>256	1.5	0.75
P167S	>256	0.375	12
D240G	>256	1	1.5
P167S/D240G	>256	0.19	0.75

7.3 Results

7.3.1 Ceftazidime resistance levels of P167S/D240G double mutant are reduced compared with single mutants.

The P167S and D240G substitutions have been observed in multiple CTX-M β-lactamase variants and are associated with 10-fold increased ceftazidime hydrolysis [458,460]. Further, introduction of the P167S and D240G substitutions into the CTX-M-3 β-lactamase results in lower ceftazidime resistance than either of the single mutants [466]. We extended these findings to the CTX-M-14 model system by determining minimum inhibitory concentrations (MICs)7.1 for ceftazidime, cefotaxime, and cephalothin for Escherichia coli harboring WT and the mutants (Table 7.1). The results show that the P167S and D240G individual substitutions both result in increased resistance to ceftazidime, whereas the P167S/D240G double mutant exhibits a loss of ceftazidime resistance compared with either the P167S or D240G single mutants (Table 7.1) [460, 466]. These data confirm the apparent incompatibility of the P167S and D240G substitutions as first suggested by Novais et al. [466] and extends the findings to the CTX-M-14 enzyme background.

Table 7.2: Enzyme kinetic parameters of CTX-M-14 β-lactamase and mutant enzymes

Enzyme	Parameter	Cephalothin	Cefotaxime	Ceftazidime
CTX-M-14	k_{cat} (s^{-1})	1400 ± 38	161 ± 9	NDa
	K_m (μM)	83 ± 6	60 ± 7	>500
	k_{cat}/K_m ($\mu M^{-1}s-1$)	17.0 ± 0.7	2.71 ± 0.16	0.0011 ± 0.00007
P167Sb	k_{cat} (s^{-1})	681 ± 36.8	297 ± 29.7	ND
	K_m (μM)	32 ± 0.8	37 ± 6.3	>500
	k_{cat}/K_m ($\mu M^{-1}s-1$)	21.1 ± 1.3	8.0 ± 1.6	0.011 ± 0.0002
D240Gb	k_{cat} (s^{-1})	471 ± 10.9	321 ± 46.2	ND
	K_m (μM)	47.1 ± 41.7	52 ± 8.6	>500
	k_{cat}/K_m ($\mu M^{-1}s-1$)	10.0 ± 2.5	6.2 ± 1.4	0.013 ± 0.0007
P167S/D240G	k_{cat} (s^{-1})	165 ± 11	139 ± 3	ND
	K_m (μM)	15 ± 3	42 ± 0.5	>500
	k_{cat}/K_m ($\mu M^{-1}s-1$)	10.8 ± 1.4	3.27 ± 0.1	0.0060 ± 0.00004

7.3.2 Antibiotic hydrolysis by the P167S/D240G double mutant is reduced compared with single mutants.

Although the P167S and D240G substitutions increase the catalytic efficiency (k_{cat}/K_m) for ceftazidime hydrolysis by ~10-fold, the activity of the P167S/D240G double mutant enzyme has not been examined [460, 462, 463, 467]. Therefore, both WT and the double mutant CTX-M-14 enzymes were purified, and their kinetic parameters were determined for hydrolysis of the oxyimino-cephalosporins cefotaxime and ceftazidime, as well as cephalothin (Table 7.2).

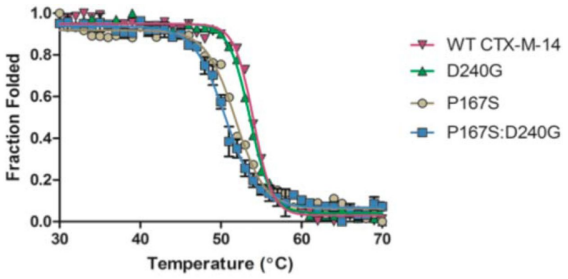
Ceftazidime hydrolysis by the WT, P167S, and D240G enzymes exhibits high K_m values (¿500 μM), which precluded determination of k_{cat} values [460]. Nevertheless, k_{cat}/K_m values for the P167S and D240G enzymes were 10-fold higher than that observed for WT CTX-M-14. If the P167S and D240G substitutions act additively, k_{cat}/K_m for ceftazidime by the double mutant should be a further 10-fold higher than that observed for the single mutants [441]. However, k_{cat}/K_m for ceftazidime hydrolysis by the double mutant was ~2-fold lower than that observed for the P167S and D240G single mutants (Table 7.2). Therefore, the P167S and D240G substitutions are antagonistic with respect to ceftazidime hydrolysis. This suggests that the presence of one substitution alters the environment of the other to perturb its contribution to catalysis [441].

The P167S and D240G substitutions were previously observed to modestly increase k_{cat}/K_m for cefotaxime hydrolysis ($\sim$2-fold) compared with the WT enzyme [460]. The P167S/D240G double mutant exhibited a k_{cat}/K_m value similar to WT and 2-fold lower than the single mutants indicating possible antagonism, as found for ceftazidime hydrolysis (Table 7.2).

The second-generation cephalosporin cephalothin is an excellent substrate for the WT CTX-M-14 enzyme (Table 7.2) [460]. The P167S and D240G substitutions reduce both k_{cat} and K_m values for cephalothin hydrolysis (Table 7.2). The P167S/D240G double mutant exhibited a further reduction in k_{cat} and K_m compared with the single mutants. Interestingly, the P167S and D240G substitutions act additively in the double mutant for cephalothin hydrolysis. Therefore, the additivity relationship between the P167S and D240G substitutions is substrate-dependent, with simple additivity observed for cephalothin and antagonism observed for ceftazidime hydrolysis, suggesting that the effects are mediated through interaction with the substrates.

7.3.3 P167S/D240G double mutant exhibits reduced stability compared with single mutants

Amino acid substitutions can affect catalysis, as shown above, but also can impact protein stability. It was previously shown that the P167S and D240G single mutants destabilize CTX-M-14 [460]. We extended this finding to the P167S/D240G enzyme using CD spectroscopy to monitor α-helix ellipticity with increasing temperature (Fig. 7.3). Previous studies showed that the WT CTX-M-14 exhibited a melting temperature (Tm) of 54.6 °C, and the single mutants D240G and P167S decreased the Tm by 0.4 and 2.8 °C, respectively [460]. The P167S/D240G enzyme exhibited a Tm of 50.5 °C, indicating that the double mutant is less stable than WT and the single mutants (Fig. 7.3).



Thermal stabilities of CTX-M variants

WT or mutant β-lactamase	T_m (°C)	ΔT_m (°C)
CTX-M-14	54.0 ± 0.09	-
D240G	53.6 ± 0.08	-0.4
P167S	51.8 ± 0.16	-2.2
P167S:D240G	50.5 ± 0.56	-3.5

Figure 7.3: Thermal stability of WT and mutant β-lactamases, as measured by CD. Mean ellipsivity is normalized and fit to a Boltzmann sigmoidal function. Tm, as determined by the Boltzmann equation, is also plotted, as is ΔTm, which is the change from the WT Tm. The Tm indicates that each single mutant is less stable than the WT CTX-M-14 enzyme, and this instability has an additive effect in the double mutant, P167S/D240G CTX-M-14. The data for CTX-M-14, D240G, and P167S are from Patel et al. [460].

7.3.4 Steady-state levels of the P167S/D240G enzyme in E. coli are reduced compared with single mutants.

The level of antibiotic resistance conferred to bacteria by a β-lactamase depends on the rate of hydrolysis, as well as the steady-state levels of enzyme expression [468]. A correlation has been shown between β-lactamase stability and expression levels in E. coli caused by increased proteolysis and aggregation of unstable proteins [468–470]. Because the P167S and D240G substitutions decrease enzyme stability and the double mutant decreases stability further, we hypothesized that the double mutant would display lower expression levels. Immunoblot analysis of whole cell lysates using α-CTX-M-14 β-lactamase polyclonal antibody showed that the P167S mutant did not significantly decrease expression levels relative to WT, consistent with previous studies (Fig. 7.4) [460]. The D240G mutant, which shows only a 0.4 °C decrease in stability relative to WT, displayed lower expression levels. Thus, although D240G has higher thermal stability than P167S, it displays lower expression levels, indicating that thermal stability does not fully correlate with expression levels. However, the P167S/D240G enzyme exhibited lower expression levels than either WT or the P167S and D240G single mu-

tants, consistent with the lower thermal stability of this mutant. Taken together, these findings provide a rationale for why the P167S/D240G double mutant has not been observed in resistant clinical isolates in that it is compromised for catalysis, stability, and expression levels compared with the P167S and D240G single mutants.

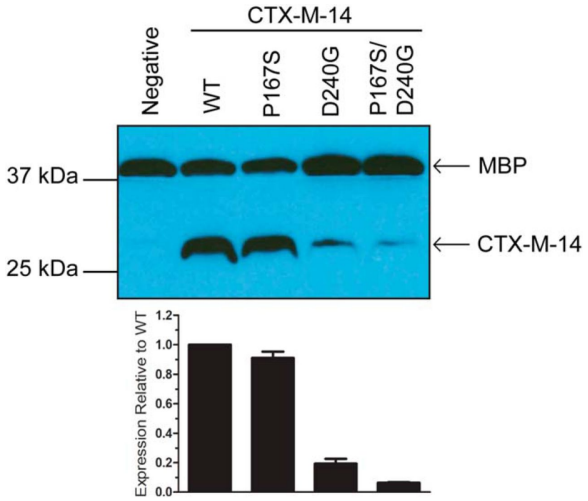


Figure 7.4: Steady-state protein levels of WT CTX-M-14 and mutant β-lactamases. Western blotting analysis with an anti-CTX-M-14 polyclonal antibody shows protein expression levels of WT CTX-M-14 and the P167S, D240G, and P167S/D240G mutants in the periplasm of recombinant E. coli cells. Analysis with a polyclonal antibody to the periplasmic protein MBP was used as a loading control. The hybridization signal was visualized by chemiluminescence and quantified by densitometry. The signal for CTX-M-14 β-lactamase was normalized to that for MBP in the same sample. The protein levels of mutant CTX-M-14 β-lactamase are expressed relative to that of WT CTX-M-14 β-lactamase in the bar graph. The quantification data in the bar graph are the averages of two independent experiments, and one representative immunoblot result is shown above the bar graph.

7.3.5 X-ray structures of P167S/D240G apo, E166A/D240G/CAZ, and E166A/P167S/D240G/CAZ acyl-enzyme complexes reveal alternate conformations of the Ω-loop.

We previously determined the X-ray structure of the P167S enzyme, which had a very similar overall structure as WT [461]. The Ω-loop, which forms the bottom of the active site, was in a folded, closed conformation with the peptide bond preceding Ser167 in a cis configuration (Fig. 7.5, A–C). The structure of the D240G enzyme was previously determined, and it also

is highly similar to the WT structure [458] (Fig. 7.5D). We next determined the structure of the P167S/D240G enzyme, which exhibits lower ceftazidime hydrolysis than either of the single mutants. The structure includes a boronic acid from the crystallization buffer in complex with Ser70 and is very similar to the WT, P167S, and D240G structures, with the Ser167 peptide bond in the cis configuration and the Ω-loop in a folded, closed conformation (Fig. 7.5, E–H, and Table E.1). A difference was noted, however, in the B-factors in the active-site 103–106 loop, suggesting increased disorder. B-factors reflect the degree to which electron density is scattered and therefore indicate how ordered an atom is in the structure [471]. The B-factors for residues in the 103–106 loop and the 164–179 Ω-loop were normalized to the overall B-factor of each structure to facilitate comparison across structures (Fig. 7.6). The normalized B-factors for the P167S/D240G structure for residues Val103 and Asn104 were higher than in the WT, P167S, and D240G structures. These findings suggest increased disorder for residues 103–104 in the P167S/D240G structure. We have previously shown that Asn104 is important for cefotaxime and ceftazidime hydrolysis, and therefore increased disorder of this residue in the P167S/D240G enzyme could result in the observed lower activity for ceftazidime hydrolysis [472].

We next determined the structures of the mutant enzymes in complex with ceftazidime to evaluate whether the presence of substrate influences active-site structure (Table E.1). The E166A mutation blocks deacylation and allows for crystallization of the acyl-enzyme complex [465]. The previously determined structure of the acyl-enzyme complex of the CTX-M-14 pseudo WT E166A enzyme with ceftazidime (E166A/CAZ) shows the Pro167 peptide bond in the cis configuration and the Ω-loop in the folded, closed conformation (Fig. 7.7A) [461]. Contacts between ceftazidime and the enzyme include hydrogen bonds between the side chains of Asn132 and Asn104 with the carbonyl oxygen of the acylamide of the ceftazidime R-2 group, as well as hydrogen bonds between the hydroxyls of Thr235 and Ser237 with the C4 carboxylate of the dihydrothiazine ring (Figs. 7.2C and 7.7A). The imino group of ceftazidime is pointed to solvent and does not interact with the enzyme. The previously determined structure

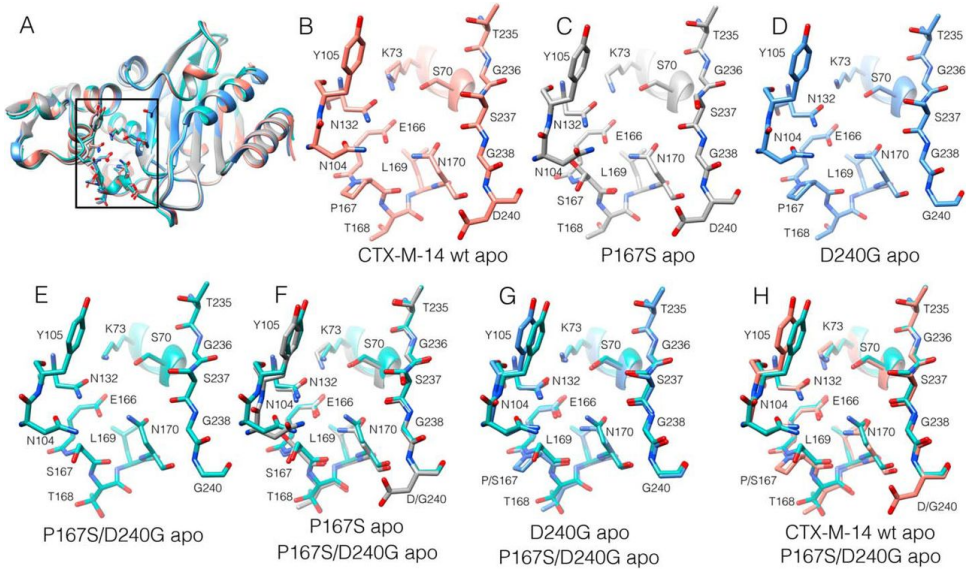


Figure 7.5: Structures of the active-site region of WT CTX-M14 β-lactamase (PDB code 1YLT), as well as P167S (PDB code 5TWD), D240G (PDB code 1YLP), and P167S/D240G mutant enzymes in the apo form. A, ribbon diagram showing structural alignment of the CTX-M-14 WT (salmon), P167S (gray), D240G (blue), and P167S/D240G (cyan) enzymes. The active-site region shown in B–H is boxed. B–E, active-site regions of CTX-M-14 WT (B), P167S (C), D240G (D), and P167S/D240G (E). Note that the P167S/D240G structure has a boronic acid in complex with Ser70, which is not shown for clarity. F, structural alignment of active-site region of P167S apo enzyme (gray) with the P167S/D240G double mutant (cyan). G, structural alignment of D240G apo enzyme (blue) with P167S/D240G (cyan). H, structural alignment of CTX-M-14 WT apo enzyme with P167S/D240G (cyan). In all panels, oxygen is shown in red, and nitrogen is in blue.

of E166A/P167S/CAZ (Fig. 7B) shows the Ser167 peptide bond in the trans configuration and the Ω-loop in an unraveled, open conformation, which widens the floor of the active site by $\sim$5 Å to accommodate ceftazidime [461]. This leads to a change in conformation of ceftazidime in the acyl-enzyme with the aminothiazole ring assuming a buried position (Fig. 7.7B) [461]. In addition, there are hydrogen bonds between the C4 carboxylate of the dihydrothiazine ring and the side chains of Thr235 and Ser237, as well as between the side chains of Asn132 and Asn104 with the carbonyl oxygen of the acylamide group (Fig. 7.7B). Further, there is a hydrogen bond between Asn104 and the carboxyl group of the imino side chain of ceftazidime. These interactions are consistent with tighter binding of ceftazidime and enhanced catalysis [461]. In addition, the normalized B-factors of Val103 and Asn104 are not increased relative to WT CTX-M-14, suggesting that the Asn104 residue is well-ordered for interaction with ceftazidime

(Fig. 7.6A). Residues 168–170, however, show elevated B-factors, suggesting that the Ω-loop has increased flexibility, consistent with its unfolded structure (Fig. 7.6B).

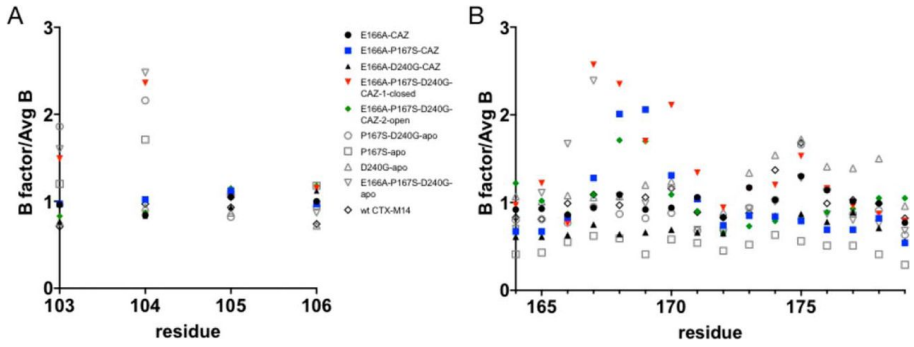


Figure 7.6: Normalized B-factors for the 103–106 loop and the 164–179 Ω-loop in the CTX-M enzyme structures. The B-factors were normalized by dividing the B-factor for a given residue by the average B-factor for the entire structure. Thus, a normalized B-factor of 1 means the B-factor at that residue is the same as the average B-factor of the structure. A, normalized B-factors for the 103–106 loop. B, normalized B-factors for the 164–179 Ω-loop. Black circle, E166A/CAZ; blue square, E166A/P167S/CAZ; black triangle, E166A/D240G/CAZ; inverted red triangle, E166A/P167S/D240G/CAZ-1; green diamond, E166A/P167S/D240G/CAZ-2; open circle, P167S/D240G apo; open square, P167S apo; open triangle, D240G apo; inverted open triangle, E166A/P167S/D240G apo; open diamond, CTX-M-14 WT.

The D240G substitution is also associated with increased ceftazidime hydrolysis [462, 464]. We therefore determined the structure of the E166A/D240G enzyme in complex with ceftazidime for comparison with the E166A and E166A/P167S acyl-enzyme structures. It was found that the peptide bond preceding Pro167 is in the cis configuration, and the Ω-loop is in the folded, closed conformation similar to the D240G apo enzyme structure and the E166A structure with ceftazidime (Fig. 7.7C). In contrast to the E166A/CAZ structure, however, the E166A/D240G/CAZ structure has the side chain of Ser237 rotated away from the C4 carboxylate and instead forms hydrogen bonds to the carboxylate of the imino side chain, which may facilitate substrate binding and catalysis (Figs. 7.2C and 7.7C).

The P167S/D240G enzyme displays lower catalytic activity toward ceftazidime than either single mutant (Table 7.2). To better understand the basis of this antagonism, we determined the structure of the E166A/P167S/D240G enzyme in complex with ceftazidime. Two structures from different space groups were obtained, and interestingly, they show different conformations of ceftazidime and the Ω-loop. In the first structure, the peptide bond preceding Ser167

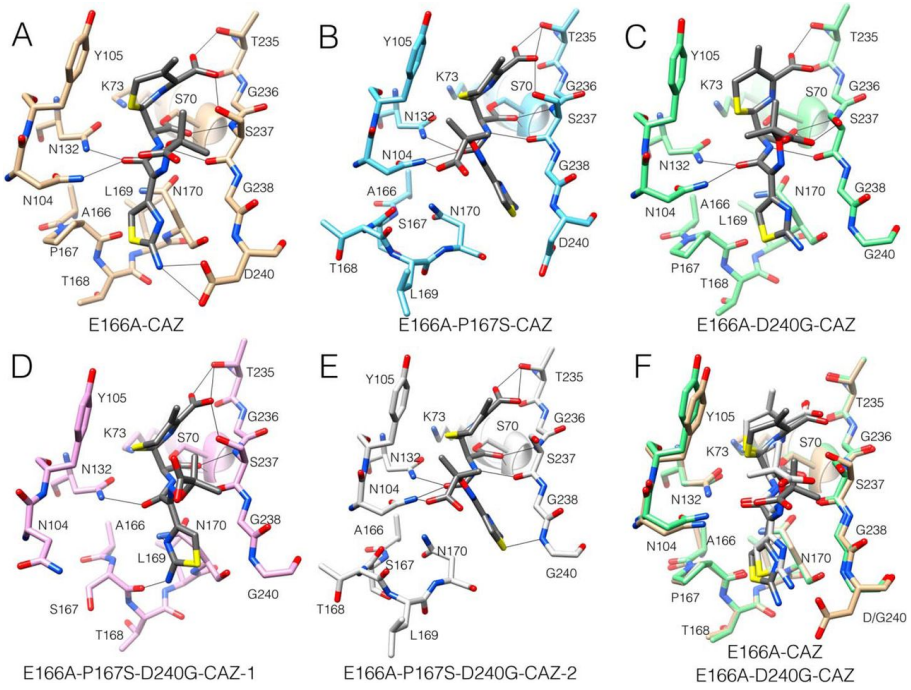


Figure 7.7: Structures of the active-site region of CTX-M-14 mutant β-lactamase acyl-enzyme complexes with ceftazidime. A, structure of the E166A mutant (tan) with ceftazidime (dark gray) trapped in the acyl-enzyme form (PDB code 5U53). Oxygen is shown in red, and nitrogen is in blue. Hydrogen bonds are indicated by thin black lines. Active-site residues are labeled. B, structure of the E166A/P167S/CAZ acyl-enzyme (light blue) (PDB code 5TW6). C, structure of the E166A/D240G/CAZ acyl-enzyme (light green). D, structure of the E166A/P167S/D240G/CAZ-1 acyl-enzyme (pink). E, structure of the E166A/P167S/D240G/CAZ-2 acyl-enzyme (white). F, structural alignment of the E166A/CAZ (tan) and E166A/D240G/CAZ (green) acyl-enzyme complexes. The ceftazidime from the E166A/CAZ structure is shown in dark gray, and that from E166A/D240G/CAZ is shown in white. The Ω-loop region remains folded in the closed form and the ceftazidime occupies a similar position with the aminothiazole ring surface exposed in these structures.

is in the trans configuration, which is in contrast to the cis bond found in the P167S/D240G apo structure (Fig. 7.5, D and E). However, the Ω-loop in the E166A/P167S/D240G/CAZ-1 structure remains in the folded, closed conformation with ceftazidime located in a similar position as that in the E166A/D240G/CAZ structure (Fig. 7.7D). There are differences, however, between these structures. First, the carboxylate group of the imino side chain of ceftazidime in the E166A/P167S/D240G/CAZ-1 structure does not contact the enzyme, in contrast to the E166A/D240G/CAZ structure (Fig. 7.7, C and D). More importantly, the positioning of the active-site 103–106 loop is altered, and the side chain of Asn104 is shifted out of the active site in the E166A/P167S/D240G/CAZ-1 structure (Fig. 7.7D). In addition, the normal-

ized B-factors for residues Val103 and Asn104 are elevated compared with the E166A/CAZ, E166A/P167S/CAZ, and E166A/D240G/CAZ structures, suggesting that Val103 and Asn104 are disordered (Fig. 7.6A). We have previously shown that the hydrogen bond between Asn104 and the acyl-amide of cefotaxime and ceftazidime is important, and a N104A mutant exhibits 10-fold lower k_{cat}/K_m for both substrates [472]. These observations suggest that the conformation of the enzyme and ceftazidime observed in the E166A/P167S/D240G/CAZ-1 structure is not consistent with hydrolysis.

It is noteworthy that the E166A/P167S/D240G/CAZ-1 structure described above was obtained by soaking a crystal with ceftazidime. Another crystal was also soaked, and the structure was determined with the same space group, but ceftazidime was not present in the active site. Interestingly, this apo structure is very similar to the structure with bound ceftazidime. The peptide bond preceding Ser167 is in the trans configuration and the Ω-loop is in the folded, closed conformation (Fig. 7.8, C and D). In addition, the 103–106 loop is in a similar position as in the ceftazidime-bound structure with the side chain of Asn104 pointed out of the active site and with elevated B-factors for Val103 and Asn104 (Fig. 7.6A). Thus, this conformation of the enzyme, and particularly the 103–106 loop, occurs in the absence of ceftazidime, in contrast to the different conformations of the E166A/P167S apo and E166A/P167S/CAZ structures where the presence of ceftazidime is apparently required to produce the conformational change.

The second E166A/P167S/D240G/CAZ structure is superimposable with that of the E166A/P167S/CAZ structure where the peptide bond preceding Ser167 is in the trans configuration, and the Ω-loop is in an unfolded, open conformation to accommodate ceftazidime (Fig. 7.7E). Because the P167S enzyme exhibits enhanced ceftazidime hydrolysis, these findings suggest that the conformation of the enzyme in the E166A/P167S/D240G/CAZ-2 structure is competent to hydrolyze ceftazidime.

The two structures of E166A/P167S/D240G/CAZ with distinct conformations of the enzyme and ceftazidime suggests there are at least two conformational substates of the P167S/D240G

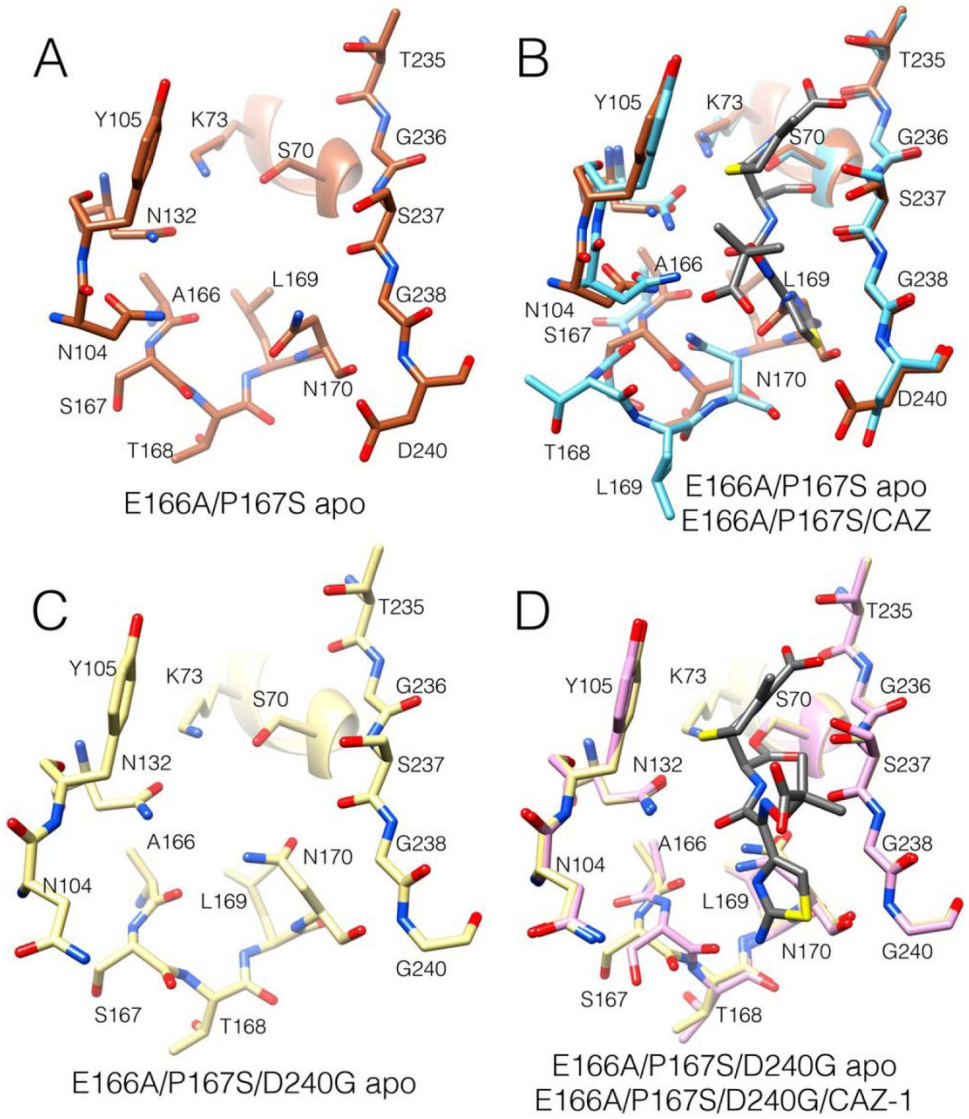


Figure 7.8: Structures of the active-site region of CTX-M-14 mutant β-lactamase acyl-enzyme complexes with ceftazidime. A, structure of the E166A mutant (tan) with ceftazidime (dark gray) trapped in the acyl-enzyme form (PDB code 5U53). Oxygen is shown in red, and nitrogen is in blue. Hydrogen bonds are indicated by thin black lines. Active-site residues are labeled. B, structure of the E166A/P167S/CAZ acyl-enzyme (light blue) (PDB code 5TW6). C, structure of the E166A/D240G/CAZ acyl-enzyme (light green). D, structure of the E166A/P167S/D240G/CAZ-1 acyl-enzyme (pink). E, structure of the E166A/P167S/D240G/CAZ-2 acyl-enzyme (white). F, structural alignment of the E166A/CAZ (tan) and E166A/D240G/CAZ (green) acyl-enzyme complexes. The ceftazidime from the E166A/CAZ structure is shown in dark gray, and that from E166A/D240G/CAZ is shown in white. The Ω-loop region remains folded in the closed form and the ceftazidime occupies a similar position with the aminothiazole ring surface exposed in these structures.

enzyme in the presence of ceftazidime. We suggest that the form with the closed Ω-loop and altered 103–106 loop with Asn104 pointed out of the active site does not efficiently hydrolyze

ceftazidime, whereas the form with the open Ω-loop is catalytically competent.

7.3.6 Molecular dynamics simulations reveal that conformational heterogeneity of the Ω-loop is greater in the single mutants than in the WT or double mutant.

To directly probe the conformational heterogeneity of the Ω-loop and acyl-enzyme complex, we conducted molecular dynamics simulations of the acylated forms of WT, D240G, P167S, and P167S/D240G. In addition to providing atomically detailed models of the distribution of structures that CTX-M adopts, the fact that no chemical reactions occur in these simulations allowed us to include Glu166 and interrogate its interactions with ceftazidime and CTX-M. Simulations of WT were initiated from a crystal structure of the acyl-enzyme complex (PDB code 5U53) [461], and simulations of P167S were initiated based on a previous model of the E166A/P167S/CAZ structure (PDB code 5TW6) [461]. Simulations of D240G were initiated from the E166A/D240G/CAZ crystal structure presented in this work. The closed-conformation crystal structure of E166A/P167S/D240G/CAZ-1 was the initial starting structure for simulations of P167S/D240G/CAZ. In all structures, Ala166 was mutated back to a glutamic acid, and a total of 2.5 μs of simulation was run for each variant.

The distribution of Ω-loop conformations observed in our simulations suggests a correlation between Ω-loop opening and ceftazidime hydrolysis activity. The WT and P167S/D240G variant with acylated ceftazidime both favor a well-defined closed conformation (Fig. 7.9). However, we note that the P167S/D240G variant with acylated ceftazidime sparsely samples open conformations of the Ω-loop, some of which are very similar to the crystallographic structure capturing the open state (Fig. 7.7, D and E, and Fig. E.3). A previous combination of simulations and experiments have also demonstrated that the WT has a sparsely populated state with an open Ω-loop [49]. In contrast, the P167S and D240G substitutions dramatically increase the probability of a diversity of open conformations. The conformational heterogeneity of P167S is

consistent with the open structure and elevated B-factors observed in the E166A/P167S/CAZ crystal structure (Figs. 7.6 and 7.7B). Although D240G also displays substantial conformational heterogeneity, it has a deeper minima for the closed state than P167S, potentially explaining why only the closed state of D240G has been observed crystallographically so far (Fig. 7.7C).

These simulations suggest that the closed conformation inhibits catalysis by favoring a conformation of Glu166 that is incompatible with a deacylation reaction, whereas the open conformation of the Ω-loop allows Glu166 to adopt a wider range of conformations, at least some of which are compatible with the requirements for deacylation. Previous work has established that Glu166 coordinates a water that plays a role in the deacylation reaction [473] and that mutation of Glu166 traps the acyl intermediate by inhibiting deacylation [465]. Examining closed structures preferentially adopted by WT and P167S/D240G reveals that the carboxyl group of Glu166 tends to hydrogen bond with Asn170, trapping Glu166 under the Ω-loop and preventing it from coordinating the water required for deacylation (Figs. 7.9, B and C, and 7.10). In the open conformations preferentially sampled by the D240G and P167S variants, the hydrogen bond between Asn170 and Glu166 is disrupted. This open conformation is stabilized by a rearrangement of the hydrogen-bonding network in the active site where Asn132 hydrogen bonds with Glu166 (Figs. 7.9C and 7.10).

The opening and closing of the Ω-loop is also associated with a rearrangement of ceftazidime in the acyl-enzyme complex (Fig. 7.10 and Fig. E.1). One major feature observed in both crystal structures and simulations is that the aminothiazole ring of ceftazidime is buried under the Ω-loop when it is open (Figs. 7.7, B and E, and 7.10). Consistent with the crystal structures of the single-mutant variants, this rearrangement of ceftazidime in the acyl-enzyme complex is facilitated by new interactions with Asn104 and the $\beta 3$ loop. In the D240G and P167S constructs, Ser237 and Asn104 form interactions with ceftazidime (Fig. 7.10 and Figs. E.1 and E.2). We also note an additional interaction between the imino group of ceftazidime and Ser237 that is present in the more open, active variants but not the more closed, inactive variants (Fig.

E.2). These interactions with ceftazidime are rare in the WT and P167S/D240G simulations (Figs. E.1 and E.2). Furthermore, Asn104 appears to point outward toward the solvent in the closed configuration (Fig. 7.10), similar to what is seen in the crystal structure of the closed configuration of the P167S/D240G/CAZ variant (Fig. 7.7D).

Overall, our simulations suggest that the CTX-M acyl-enzyme complex is in equilibrium between inactive and active conformations and that the P167S and D240G variants have a higher probability of adopting active conformations (Fig. 7.10). In the inactive conformation the Ω-loop is closed, burying the Asn170–Glu166 hydrogen bond under the aminothiazole ring (Fig. 7.10). Opening of the Ω-loop and rearrangement of ceftazidime via burial of the aminothiazole ring and coordination between the imino group and Asn104 and the β3 loop likely transitions CTX-M into a catalytically competent state. These rearrangements of the Ω-loop and ceftazidime allow Glu166 to coordinate a water molecule that can access the ester bond of the ceftazidime–acyl-enzyme complex, facilitating the deacylation reaction. Taken together, the crystallography and molecular dynamics results indicate that the P167S and D240G substitutions promote an open conformation of the Ω-loop that creates access for ceftazidime and allows Glu166 to sample conformations consistent with deacylation, whereas the WT and P167S/D240G mutant exhibit a closed Ω-loop conformation that constrains access for ceftazidime and prevents Glu166 from efficiently coordinating water for deacylation.

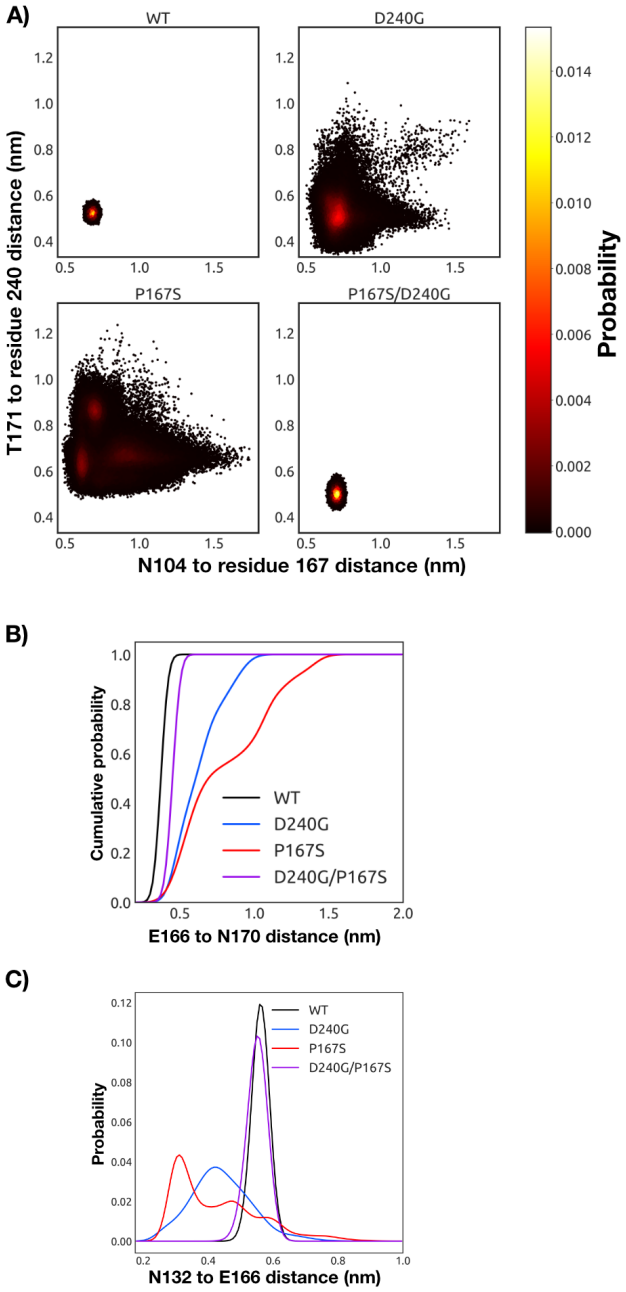


Figure 7.9: The conformational heterogeneity of the Ω-loop is greater in the single mutants than in the WT or double mutant. A, joint distributions of two Cα–Cα distances that capture the conformational heterogeneity of the Ω-Loop: (i) Asn104 to position 167 on one side of the Ω-loop and (ii) Thr171 to position 240 on the other side. Distributions are shown for WT (top left panel), D240G (top right panel), P167S (bottom left panel), and P167S/D240G (bottom right panel). Each point represents a snapshot from the molecular dynamics simulations colored according to its probability based on a 2D histogram. B, cumulative distribution of distances between the Cγ atom of Glu166 and the Nδ atom of Asn170 for WT (black), D240G (blue), P167S (red), and P167S/D240G (purple), capturing the loss of interaction between Glu166 and Asn170 in the open state. C, distribution of distances between the sidechain Nδ atom of Asn132 and the Cγ of Glu166 for WT (black), D240G (blue), P167S (red), and P167S/D240G (purple), capturing the newly formed interaction between Glu166 and Asn132 in the open state.

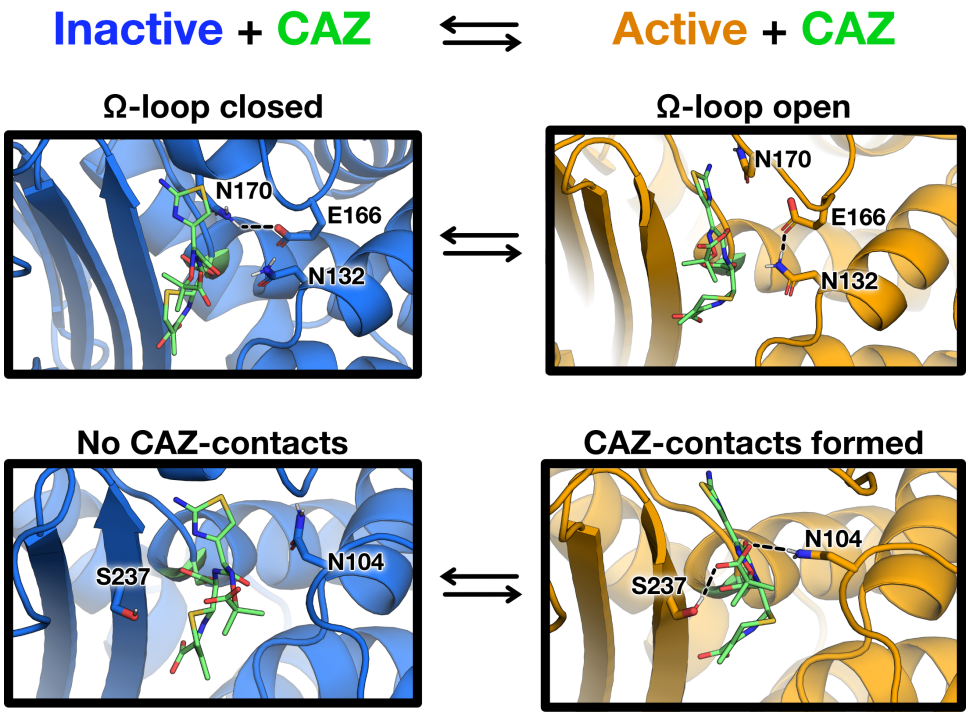


Figure 7.10: Conformational changes between inactive and active forms of the acyl-enzyme. Representative structures of the CTX-M acyl-enzyme complex with ceftazidime (labeled CAZ, green) highlighting inactive (blue, left panels) and active (orange, right panels) conformations. Conformations of the Ω-loop (top panels) and residues contacting ceftazidime (bottom panels) are shown, with relevant residues labeled. Key hydrogen bond interactions are depicted in dashed lines (black).

7.4 Discussion

The CTX-M β-lactamases emerged in the late 1980s and are characterized by their ability to efficiently hydrolyze cephalosporins, particularly the oxyimino-cephalosporin cefotaxime [456,457]. For example, the catalytic efficiency for cefotaxime hydrolysis by CTX-M enzymes is 1500-fold higher than that exhibited by the common TEM-1 β-lactamase [453]. Nevertheless, the related oxyimino-cephalosporin ceftazidime is poorly hydrolyzed by CTX-M enzymes with k_{cat}/K_m values 1000-fold lower than those observed for cefotaxime [453]. Natural variants of the TEM-1 enzyme have evolved through mutations that provide increased rates of ceftazidime catalysis. Similarly, the P167S and D240G substitutions have been found in multiple variants of CTX-M enzymes that exhibit increased ceftazidime hydrolysis [457,462–464,474]. Each of these substitutions results in a 10-fold increased k_{cat}/K_m value for ceftazidime hydrolysis [458,460]. However, variants containing both substitutions have not been observed, despite the prediction that such variants would exhibit increased hydrolysis. Here we have shown that the failure of the double mutant to emerge in natural isolates is due to epistasis resulting from decreased stability and lower bacterial expression levels of the P167S/D240G enzyme, as well as antagonism between the substitutions with respect to catalysis.

The combination of amino acid substitutions that each increase catalytic activity can display simple additivity or cooperativity when introduced together into an enzyme [441]. For additive interactions, the fold change of the double mutant is expected to be the product of the fold changes of the single mutants. Such additive combinations indicate that the substitutions have independent effects on catalysis [441]. However, not all substitutions act additively. The CTX-M P167S and D240G substitutions are antagonizing in the double mutant (Table 7.2). This negative cooperativity suggests that the substitutions interact, either directly or indirectly, and that the interaction has a negative effect on ceftazidime catalysis [441]. In the case of the P167S and D240G substitutions, the interaction is not via direct contact, because the α-carbons are located 10.7 Å apart.

X-ray crystallography and molecular dynamics simulations of the P167S, D240G, and P167S/D240G mutants provide a rationale for the increased ceftazidime hydrolysis by the single mutants and the negative cooperativity observed for the double mutant. The acyl-enzyme complex of the E166A/P167S/CAZ X-ray structure shows a trans configuration peptide bond preceding residue 167 and an unfolded Ω-loop in an open conformation with the aminothiazole ring of the antibiotic in a buried position. This interaction increases van der Waals contacts and hydrogen bonds between the enzyme and ceftazidime and is consistent with enhanced catalytic efficiency toward ceftazidime. In contrast, the E166A/D240G/CAZ structure revealed a closed conformation. Two structures were obtained for the E166A/P167S/D240G ceftazidime acyl-enzyme, one of which is superimposable with the open Ω-loop structure of E166A/P167S/CAZ and, based on this similarity, is proposed to hydrolyze ceftazidime. The second structure, however, displays a closed Ω-loop and an altered conformation of the 103–106 loop such that the critical Asn104 residue is turned out of the active site and does not interact with ceftazidime, suggesting reduced ceftazidime catalysis.

Although the E166A/P167S/CAZ and E166A/D240G/CAZ structures suggest that the substitutions act by different mechanisms, 2.5-μs molecular dynamics simulations of each variant indicate that both the P167S and D240G substitutions promote an open conformation of the Ω-loop to accommodate ceftazidime and that the WT and P167S/D240G enzymes exhibit reduced ceftazidime hydrolysis because open conformations of the Ω-loop are less probable.

The TEM-1 β-lactamase is $\sim$35% identical in amino acid sequence to CTX-M enzymes and efficiently hydrolyzes penicillins and many cephalosporins, but oxyimino-cephalosporins are poor substrates [453]. Nevertheless, natural variants of TEM-1 have evolved that exhibit increased catalytic efficiency for cefotaxime and ceftazidime hydrolysis [475, 476]. These variants, termed extended-spectrum β-lactamases (ESBLs), contain 1–5 amino acid substitutions, and multiply substituted enzymes are common. Many of the substitutions found in these variants act additively when combined [453]. Some combinations of mutations, however, are not additive. A well-studied example is the combination of the R164S and G238S substitutions.

The Gly238 residue is on the β3 strand (analogous to Gly238 in CTX-M-14), and Arg164 is at the base of the Ω-loop. Each of these substitutions results in increased enzyme activity toward cefotaxime and ceftazidime, yet the double mutant has reduced activity [477, 478]. Dellus-Gur et al. [477] determined the structure of the TEM-1 G238S and R164S enzymes. The G238S-containing enzyme exhibited two dominant conformations of the G238 loop, whereas the R164S substitution induced an ensemble of conformations of the Ω-loop. The structure of the R164S/G238S double mutant, however, exhibited a wider ensemble of conformations of the Ω-loop than the single mutants [477]. Based on these results, it was hypothesized that the entropic cost of the substrate selecting suitable conformations among many alternatives results in the low activity for cefotaxime hydrolysis by the double mutant, accounting for the negative epistasis observed for the combination [477].

In the case of the negative epistasis observed with the P167S and D240G combination in CTX-M β-lactamase, multiple conformations of the enzyme also appear to play a role. Based on our molecular dynamics simulation results, the P167S and D240G substitutions are analogous to the R164S substitution in TEM-1, where the substitutions induce an ensemble of conformations, some of which are predicted to be capable of hydrolyzing ceftazidime. The CTX-M P167S and D240G substitutions antagonize each other in the double mutant, similar to the TEM-1 R164S and G238S substitutions. The mechanism of antagonism, however, is different, with the CTX-M P167S/D240G double mutant showing a reduced probability of sampling multiple conformations of the Ω-loop, whereas the TEM-1 R164S/G238S double mutant Ω-loop samples an excess of conformations [477]. Nevertheless, both cases demonstrate the importance of conformational heterogeneity of active-site loops in controlling catalytic activity and evolutionary trajectories.

Several studies have provided evidence that the conformation of the Ω-loop is an important determinant of substrate specificity of class A β-lactamases, particularly with regard to the hydrolysis of oxyimino-cephalosporins. As described above, the TEM ESBL mutation R164S is thought to broaden the specificity of the enzyme by increasing the conformational heterogene-

ity of the Ω-loop [477]. In addition, the structure of an apo enzyme form of a triple mutant of the TEM enzyme containing the substitutions W165Y/E166Y/P167G that hydrolyzes ceftazidime shows the Ω-loop in an unfolded, open conformation similar to that observed for the CTX-M E166A/P167S/CAZ and E166A/P167S/D240G/CAZ-2 structures [479]. Further, computational studies predict that TEM ESBL substitutions that broaden the specificity of the enzyme to include cefotaxime stabilize conformations of the Ω-loop that facilitate substrate binding [9].

The results also suggest an important role for the active-site 103–106 loop in cefotaxime and ceftazidime hydrolysis. We recently showed that an N106S mutation in the 103–106 loop that is found in CTX-M enzymes from clinical isolates lowers cefotaxime and ceftazidime hydrolysis because of a change in conformation of the loop [472]. Asn106 is at the base of the loop and not in the active site. However, the N106S substitution changes the hydrogen-bonding network connectivity in the loop such that the side chain of Asn104 rotates out of the active site, thereby eliminating a hydrogen bond with substrate. Further experiments showed that an N104A mutant exhibits 10-fold reduced catalytic efficiency for oxyimino cephalosporin hydrolysis, suggesting that the hydrogen bond is important for catalysis. Thus, the conformation of the 103–106 loop is a determinant of substrate specificity [472]. In this study, it was found that the P167S/D240G enzyme, which exhibits reduced ceftazidime hydrolysis, has increased B-factors for the 103–106 loop, suggesting disorder in Asn104 that is consistent with reduced activity. In addition, in the structure of the E166A/P167S/D240 apo enzyme the B-factors of the 103–106 loop are increased, and in one of the structures of E166A/P167S/D240G in complex with ceftazidime, the side chain of Asn104 is rotated out of the active site, again consistent with decreased ceftazidime hydrolysis. Thus, although the P167S and D240G substitutions are not in the 103–106 loop, the antagonism between the substitutions is at least partially reflected in changes in the conformation of the loop.

Thermal stability studies of the WT, P167S, D240G, and P167S/D240G enzymes show that the single mutants are less stable than WT, whereas the double mutant is less stable than either

single mutant. There is some correlation between stability and protein expression levels in that the P167S/D240G mutant is the least stable and is also expressed at the lowest levels among the mutants. However, also note the P167S mutant is less stable than D240G but is expressed at higher levels, suggesting there are exceptions to the stability/protein expression level correlation. Some recent studies have shown that lower enzyme stability correlates with increased flexibility and increased cephalosporin hydrolysis in β-lactamases [480, 481]. Here, we do not observe a correlation between stability and flexibility in that the P167S and D240G mutants readily sample multiple conformations and yet are more stable than P167S/D240G, which samples fewer conformations. Further, we do not observe a correlation between stability and catalytic activity toward ceftazidime because P167S/D240G has low stability but also low activity.

Taken together, the results presented here suggest that active-site loops play an important role in the substrate specificity and evolutionary capacity of β-lactamases. Class A β-lactamases such as TEM and CTX-M can evolve altered substrate specificity by mutations that change the conformation of active-site loops. An active site with flexible loops loosely associated with a highly ordered, stable scaffold structure has been described as fold polarity, and there is evidence that such an organization facilitates the evolution of new functions because of a tolerance to changes in the loops without drastically destabilizing the enzyme [443]. Such an organization is clearly advantageous for antibiotic resistance enzymes such as CTX-M β-lactamases that are under selective pressure for altered substrate specificity.

7.5 Methods

7.5.1 Bacterial strains and plasmids.

The CTX-M-14-pTP123 plasmid was used for site-directed mutagenesis, MIC determinations, and immunoblotting. This plasmid was constructed by inserting the blaCTX-M-14 gene into

the previously described pTP123 plasmid. CTX-M-14-pTP123 has a chloramphenicol (CMP) resistance marker and $\beta-$-lactamase expression is controlled by the isopropyl$-\beta-$-D-galactopyranoside (IPTG)–inducible P_{trc} promoter [482]. Under conditions without IPTG, protein expression is maintained at a basal level; these were the conditions under which MIC determination and immunoblotting were performed. The E. coli strain XL1-Blue (recA1 endA1 gyrA96 thi-1 hsdR17 supE44 relA1 lac [F9 proAB lacI^{q}ZM15 Tn10 (Tet^r)]) (Stratagene, Inc., La Jolla, CA) was used as the host for the construction of the P167S/D240G CTX-M-14 mutant via site-directed mutagenesis, as well as MIC determination. The E. coli strain RB791 (W3110 lacIqL8) was used as the host for the determination of protein expression levels of WT CTX-M-14 β-lactamase and its mutants [483]. For protein purification, WT CTX-M-14 and the mutants were expressed in the pET28a plasmid using the protocol outlined by Patel et al. [460]. CTX-M-14 and mutants in the pET28a plasmid were expressed in BL21(DE3) [fhuA2 (lon) ompT gal (λ DE3) (dcm) ΔhsdS λ DE3 = λsBamHIo ΔEcoRI-B int::(lacI::PlacUV5::T7 gene1) i21 Δnin5] [484].

7.5.2 Site-directed mutagenesis.

The CTX-M-14 P167S/D240G mutant was constructed using the D240G mutant in pTP123 as template for QuikChange mutagenesis using 1 unit of Phusion DNA polymerase (New England Biolabs, Ipswich, MA) and 0.4 μM P167S primer (5'-CTGGATCGCACTGAAAGCACGCTGAATACCGCC-3') [460]. Primers were obtained from Integrated DNA Technologies (Coralville, IA). Thermocycler products were digested with DpnI (New England Biolabs) and transformed into electrocompetent E. coli XL1-Blue cells and selected on LB agar supplemented with 12.5 μg/ml chloramphenicol. The DNA sequence of the resulting mutant was confirmed by DNA sequencing (Genewiz, Plainfield, NJ). The E166A/P167S/D240G mutant was constructed by QuikChange mutagenesis with a primer encoding the E166A/P167S mutations (5'-GATCGCACTGCTCCTACGCTGAAT-3') using CTX-M-14 P167S/D240G pTP123 as template and confirmed using DNA sequencing.

7.5.3 Minimum inhibitory concentration determinations.

MICs for cephalothin were determined by Etest strip (BioMérieux, Marcy-l'Étoile, France). This was performed by growing a single colony of E. coli XL1-Blue harboring the pTP123 plasmid with either WT, mutant CTX-M-14, or empty vector overnight in LB supplemented with 12.5 μg/ml CMP in a shaking incubator at 37 °C. The overnight culture was diluted 10^2 and spread onto LB agar containing CMP, an Etest strip was placed on the agar, and the MIC was determined based on the zone of inhibition.

MIC determinations for cefotaxime were performed by broth dilution. Again, a single colony of E. coli XL1-Blue harboring pTP123 with WT or mutant CTX-M-14 or empty vector was grown overnight in LB with CMP in a shaking incubator at 37 °C. The cultures were diluted 10^4, and 100 μl of culture was used to inoculate 2 ml of LB supplemented with increasing concentrations of cefotaxime in 14-ml test tubes. Concentrations of cefotaxime (in μg/ml) used for WT and D240G were 0, 1, 1.5, 2, and 3. The concentrations used for the P167S and P167S/D240G mutants were 0, 0.1875, 0.25, 0.375, and 0.5, and for pTP123 empty vector control, the concentrations were 0, 0.03, 0.045, and 0.06. The cultures were incubated with shaking for 18 h at 37 °C. The concentration at which no visible growth was observed was reported as the MIC.

MIC determinations for ceftazidime were performed by broth dilution, with a single colony of E. coli XL1-Blue harboring pTP123 with WT or mutant CTX-M-14 or empty vector being grown overnight in LB with CMP, as above. Saturated cultures were diluted 104, and 25 μl was used to inoculate 500 μl of LB supplemented with increasing concentrations of ceftazidime in a deep-well 96-well plate. Concentrations of ceftazidime (in μg/ml) used for WT, D240G, P167S/D240G, and pTP123 empty vector were 0, 0.12, 0.19, 0.25, 0.38, 0.5, 0.75, 1, 1.5, 2, 3, and 4; for P167S, the concentrations were 0, 0.38, 0.5, 0.75, 1, 1.5, 2, 4, 6, 8, and 12. The 96-well plate was covered with a sterile, breathable seal (Excel Scientific, Victorville, CA) and incubated shaking at 37 °C for 18 h. The concentration at which no growth was observed

($A_{600} < 0.1$) was recorded as the MIC.

7.5.4 Immunoblotting.

To determine the effects of the P167S/D240G mutation on steady-state protein expression, single colonies of E. coli RB791 harboring pTP123 or the recombinant pTP123 encoding WT or mutant CTX-M-14 β-lactamase were incubated overnight with shaking in 2× YT medium supplemented with 12.5 μg/ml CMP at 37 °C. A total of 10 ml of 2× YT medium with CMP was inoculated with 100 μl of overnight culture and incubated at 37 °C while shaking until the A600 reached between 0.9. The cells were pelleted, and the periplasmic proteins were extracted by osmotic shock as described previously [472]. The proteins were fractionated by SDS-PAGE and transferred onto a nitrocellulose membrane (GE Healthcare). The membrane was probed with a rabbit serum raised against CTX-M-14 protein and a rabbit serum raised against maltose-binding protein (MBP) (a gift from Dr. Anna Konovalova, University of Texas Health Science Center at Houston), which functions as a loading control. Then the membrane was probed with a donkey anti-rabbit secondary antibody conjugated with horseradish peroxidase (GE Healthcare). After development of the immunoblot with the SuperSignal West Pico chemiluminescent substrate (Thermo Fisher Scientific), the hybridization signals of CTX-M-14 β-lactamase and MBP were quantified by densitometry using ImageJ software (National Institutes of Health). The signal for WT and mutant CTX-M-14 β-lactamase was normalized to that for MBP.

7.5.5 Protein purification.

WT CTX-M-14 β-lactamase and the P167S/D240G mutant were expressed from a pET28a vector in E. coli BL21(DE3) cells. Proteins expressed in this plasmid have an N-terminal His tag with a TEV protease cleavage site. E. coli BL21(DE3) cells harboring the CTX-M-14 plas-

mid were used to inoculate LB supplemented with 25 μg/ml kanamycin and incubated at 37 °C in a shaking incubator until the A600 reached $\sim$0.9, at which time IPTG was added to yield a final concentration of 0.2 mM to induce protein expression. The culture was then grown for 20 h at 23 °C in a shaking incubator. The culture was centrifuged at 8000 rpm at 4 °C, and the pellet was stored overnight at -80 °C. The pellet was then thawed on ice and resuspended in 30 ml of lysis buffer (20 mM HEPES, pH 7.4, 300 mM NaCl, 20 mM imidazole). The cells were lysed using a French Press at 1250 p.s.i. and a probe sonicator, followed by centrifugation for 50 min at 10,000 rpm and filtration of the supernatant with a 0.45-μm filter (EMD Millipore, Billerica, MA). Filtered lysate was then bound to a HisTrap FF column (GE Healthcare) equilibrated with the lysis buffer. The CTX-M-14 enzyme was eluted using 20–500 mM imidazole gradient in the lysis buffer. Pure fractions containing His-CTX-M-14 protein were pooled, concentrated, and buffer-exchanged to the lysis buffer using 10-kDa molecular mass cutoff centrifugal filters (EMD Millipore). 0.25 mg of TEV protease was added to the His-tagged enzyme and incubated overnight at 4 °C. TEV protease and uncleaved His-CTX-M-14 protein were removed by incubation with nickel–Sepharose Hi-Performance beads (GE Healthcare). CTX-M-14 proteins were further purified by gel-filtration chromatography with Superdex 75 Increase column using (20 mM HEPES, pH 7.4, 150 mM NaCl) as running buffer. The fractions corresponding to monomer of CTX-M-14 WT or P167S/D240G mutant protein were pooled and concentrated with 10-kDa molecular mass cutoff centrifugal filters. The purity of purified proteins was $>$95% determined by SDS-PAGE. Their concentrations were determined by measuring the absorbance at 280 nm with DU800 spectrophotometer (Bec$K_m$an Coulter) and using an extinction coefficient of 23, 950 M^{-1} cm^{-1}.

7.5.6 Determination of thermal stabilities.

Thermal stabilities of the WT and mutant enzymes were determined as previously described [460]. In short, the fraction of folded protein was measured with a spectropolarimeter at 222 nm, while the temperature was increased from 30 to 70 °C at a rate of 0.01 °C/s. The melting

temperature (T_m), the temperature midpoint of protein unfolding, was determined by fitting the data to a single Boltzmann two-state model using GraphPad Prism 6 (San Diego, CA) [469].

7.5.7 Steady-state enzyme kinetic parameters.

Michaelis–Menten steady-state kinetic parameters were measured as previously described [460, 485]. The kinetic parameters for CTX-M-14 P167S and D240G are from Patel et al. [460]. Antibiotic hydrolysis was measured at 30 °C in 50 mM phosphate buffer (pH 7.0) containing 1 μg/ml BSA. BSA was added to stabilize β-lactamase when it is diluted to low concentration for kinetic assays. Cephalothin, cefotaxime, and ceftazidime hydrolysis were measured at 262, 264, and 260 nm, respectively [460]. K_m and k_{cat} were determined by fitting the initial velocities of increasing substrate concentrations to the Michaelis–Menten equation using GraphPad Prism 6. For ceftazidime, which has a $K_m > 500$ μM, k_{cat}/K_m was estimated using the equation, $v = k_{cat}/K_m[E][S]$, where $[S] \ll K_m$. All measurements were performed at least in duplicate. k_{cat} and K_m values from each run were averaged, and the standard deviations reported are the sums of the percent standard deviations of k_{cat} and K_m [460].

7.5.8 Protein crystallization and structure determination.

Crystallization conditions were screened based on previously solved crystal structures for CTX-M-14. Purified P167S/D240G enzyme in 50 mM phosphate buffer, pH 7.0, was concentrated to 40 mg/ml, and protein was mixed with mother liquor 1:1 in a 200-nl drop and grown by hanging-drop vapor diffusion. Diffraction-quality crystals were obtained in 0.1 M MIB buffer (malonic acid, imidazole, and boric acid buffer), pH 4.0, 25% (w/v) PEG 1500, and were harvested and cryoprotected in 25% glycerol: 75% mother liquor. The crystals were plunged in liquid nitrogen and sent to Beamline 5.0.2 at the Advanced Light Source (Berkeley, CA) for data collection. Because the first data set appeared to show high twinning, a second data set

was collected on the same crystal. This data set was processed at 1.5 Å in the space group P41212 using HKL200 and the Phaser program from the CCP4 suite was used for molecular replacement. CTX-M-14 (PDB code 1YLT) was used as a phasing model [458]. Refinement was performed using REFMAC5 and phenix.refine, as part of the Phenix program suite [486]. The model was built manually using COOT [487].

E166A/P167S/D240G was crystallized by concentrating the protein in 50 mM phosphate buffer, pH 7.0, to 40 mg/ml and mixing with mother liquor 1:1 in a 200-nl drop and grown by hanging-drop vapor diffusion. Crystals from the condition containing 0.1 M PCB buffer, pH 6, 25% (w/v) PEG 1500 were soaked for 24 h in 25 mM ceftazidime, 20% glycerol:80% mother liquor. Structure determination indicated an acyl-enzyme complex with ceftazidime. Crystals grown in the condition containing 0.2 M CaCl, 0.1 M sodium acetate, pH 5, 20% (w/v) PEG 6000 were also soaked for 24 h in 25 mM ceftazidime, 20% glycerol:80% mother liquor but were not in complex with ceftazidime, resulting in the apoenzyme. The data were collected on Beamline ALS 501 and was processed as described above. The E166A/P167S/D240G enzyme was also crystallized by concentrating the protein in 50 mM phosphate buffer, pH 7.0, to 36 mg/ml, mixing with mother liquor 1:1 in a 200-nl drop, and grown by hanging-drop vapor diffusion. Crystals obtained in the condition 0.1 M MMT (1:2:2 ratio of DL-malic acid:MES:Tris base), pH 6.0, 25% (w/v) PEG 1500 were soaked for 24 h in 25 mM ceftazidime, 25% glycerol:75% mother liquor. The data were collected on Beamline ALS 821. Structure determination revealed an acyl-enzyme complex with ceftazidime with the Ω-loop in an open conformation.

E166A/D240G was crystallized by concentrating the protein in 50 mM phosphate buffer, pH 7.0, to 36 mg/ml and mixing with mother liquor 1:1 in a 200-nl drop and grown by hanging-drop vapor diffusion. Crystals grown in 0.1 M Tris-HCl, pH 8.5, 25% (w/v) PEG 3000 were soaked for 24 h in 25 mM ceftazidime, 25% glycerol:75% mother liquor. The data were collected on Beamline ALS 822. However, structure determination revealed this to be the apo enzyme. Crystals grown in the condition 0.2 M NaCl, 0.1 M Tris, pH 8.0, 20% (w/v) PEG 6000 were soaked for 24 h in 15 mM ceftazidime, 25% glycerol:75% mother liquor, the data were

collected on Beamline ALS 822, and structure determination revealed an acyl-enzyme complex with ceftazidime. The data set was processed as described above for the P167S/D240G enzyme. X-ray crystallography statistics are listed in Table E.1.

7.5.9 Molecular dynamics simulations.

As described previously [9], simulations were run at 300 K with the GROMACS software package [9,145,222,488] using the Amber03 force field [146] and TIP3P explicit solvent [143]. Mutations were introduced in PyMOL [144], and parameters for the acyl group were generated with the generalized amber force field [489–491]. A total of 2.5 μ s of simulation were run for each variant.

7.6 Author contributions

C. A. B., B. V. V. P., G. R. B., and T. P. conceptualization; C. A. B., L.H., Z.S., M.P.P., S.S., J.R.P., and B.S. investigation; C. A. B., L.H., Z.S., M.P.P., S.S., J.R.P., and B.S. methodology; C. A. B. and T. P. writing-original draft; C. A. B., L.H., Z.S., M.P.P., S.S., J.R.P., B.S., B. V. V. P., G. R. B., and T. P. writing-review and editing; L.H., Z.S., S.S., J.R.P., B.S., B. V. V. P., G. R. B., and T. P. formal analysis; Z.S. data curation; S.S. visualization; B. V. V. P., G. R. B., and T. P. supervision; B. V. V. P., G. R. B., and T. P. funding acquisition.

7.7 Acknowledgments

We thank Hiram Gilbert for discussions and comments on the manuscript. The ALS-ENABLE Beamlines are supported in part by National Institutes of Health, NIGMS Grant P30 GM124169-01. The Advanced Light Source is a Department of Energy Office of Science User Facility

under Contract DE-AC02-05CH11231.

This work was supported by National Institutes of Health Grants R01 AI32956 (to T. P.) and R01 GM12400701 (to G. R. B.) and by Robert Welch Foundation Grant Q1279 (to B. V. V. P.). The authors declare that they have no conflicts of interest with the contents of this article. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

7.8 Additional information.

The atomic coordinates and structure factors (codes 6V5E, 6V6G, 6V6P, 6V7T, 6V8V, and 6V83) have been deposited in the Protein Data Bank (http://wwpdb.org/).

Table 7.3: The abbreviations used are:

Abbreviation	Expanded name
MIC	*Minimum inhibitory concentration*
PDB	*Protein data bank*
ESBL	*Extended spectrum β-lactamase*
CMP	*chloramphenicol*
IPTG	*isopropyl-β-D-galactopyranoside*
MBP	*maltose-binding protein*

Chapter 8

Conclusions

8.1 Main findings

This thesis demonstrates how to measure allosteric communication in proteins from long-timescale MD simulations, and illustrates how insight from allosteric networks can explain fundamental protein behaviors and open up new therapeutic opportunities. This work also highlights the utility of protein conformational ensembles and the tremendous scientific value encoded within protein energy landscapes. The data presented provides further support to the hypothesis that a protein's thermal fluctuations contain all of its functionally relevant states.

In **Chapter 2**, the CARDS methodology is described as a means to integrate concerted structural and disorder-mediated correlations into a holistic view of allosteric communication. Applications of CARDS to the Catabolite Activator Protein (CAP), shown experimentally to undergo allosteric coupling via conformational entropy, are described. Specifically, we showed that examining the coupling of every residue to a known cAMP binding site naturally highlights regions of the protein that are known to be impacted by cAMP binding. Decomposing correlations into disorder-mediated and purely structural components demonstrates an important role for disorder-mediated coupling in the absence of concerted structural changes. Our

global communication metric also provides a means to identify important functional sites without foreknowledge of their existence and locations. I expect CARDS to be of great utility for understanding allostery in systems where it is already known to occur, as well as for predicting allostery in systems where it has yet to be observed.

In **Chapter 3**, we describe the complete process of G protein activation and GDP release, specifically focusing on Gαq, which has the slowest dissociation rate. These results reveal a previously unobserved intermediate that defines the rate-limiting step for GDP release and, ultimately, G protein activation. The model synthesizes a wealth of experimental data and previous analyses, and builds upon decades of literature to create a single, unified model. We highlight roles in the allosteric network for regions of known importance and identify new regions of value in the allosteric network coupling the receptor- and nucleotide-binding sites. The consistency of our model with a wide variety of structural and biochemical data suggests that it is a promising foundation for future efforts to understand the determinants of GPCR-Gα interaction specificity, how mutations cause aberrant signaling and disease, and how small molecule inhibitors modulate Gα activation. The model also adds weight to the growing appreciation for the fact that a protein's spontaneous fluctuations encode considerable information about its functional dynamics. I hope this approach and combination of analyses will prove valuable for understanding other slow conformational changes and unbinding processes.

Described in **Chapter 4** is a cryptic allosteric site in the IID of the Ebola VP35 protein that provides a new therapeutic opportunity against this essential viral protein. We used adaptive sampling simulations to access more of the ensemble of conformations that VP35 adopts, uncovering an unanticipated cryptic pocket. Application of CARDS to these simulations suggested that the cryptic pocket is allosterically coupled to the blunt end-binding interface and, therefore, could modulate biologically-important interactions. Subsequent experiments highlighted that fluctuations within the folded state of the IID expose two buried cysteines that line the proposed cryptic pocket to solvent. Moreover, covalently modifying these cysteines to stabilize the open form of the cryptic pocket allosterically disrupts binding to dsRNA blunt ends

by at least 5-fold. Therefore, it may be possible to attenuate the impact of viral replication and restrict pathogenicity by designing small molecules to target the cryptic allosteric site we report here. More generally, these results speak to the power of simulations to provide simultaneous access to both hidden conformations and dynamics with atomic resolution. Thus I hope this demonstrate the potential of simulations as a means to uncover unanticipated features of proteins' conformational ensembles, such as cryptic pockets and allostery, providing a foundation for the design of further experiments. We anticipate such simulations will enable the discovery of cryptic pockets and cryptic allosteric sites in other proteins, particularly those that are currently considered undruggable.

Chapter 5 and 6 highlight the power of Folding@home to tackle a myriad of biological problems, both fundamental and with potential relevance to therapeutic design. Work in **Chapter 5** revealed that SARS-CoV-2 N protein undergoes phase separation with RNA when reconstituted in vitro. In this work we propose a model where a single-genome condensate forms through N protein gRNA interaction, driven by a small number of high-affinity sites. This (meta)-stable single-genome condensate undergoes subsequent maturation, leading to virion assembly. In this model, condensate-associated N proteins are in exchange with a bulk pool of soluble N protein, such that the interactions that drive compaction are heterogeneous and dynamic. The model provides a physical mechanism in good empirical agreement with data for N protein oligomerization and assembly.

In **Chapter 6** we describe how the excascale power of Folding@home allows us to hunt for druggable opportunities throughout the entire SARS-CoV-2 proteome. The pandemic caused by SARS-CoV-2 necessitated a call-to-arms; a call that over a million citizen-scientists answered to generate 0.1 seconds of simulation data. We find that spike proteins have a strong trade-off between making ACE2 binding interfaces accessible to infiltrate cells and conformationally masking epitopes to subvert immune responses. These simulations also provide an atomically detailed roadmap for targeting proteins for vaccines and antivirals. We describe a number of cryptic pockets that we identify throughout the proteome of SARS-CoV-2, with

more to be described as they are discovered. For each protein system in Table 6.1, an extraordinary amount of sampling has led to the generation of a quantitative map of its conformational landscape.

Finally, in **Chapter 7** we demonstrate how molecular simulations, when integrated with experiments, can be used to study the biophysical impact of mutations, such as those that grant Ceftazidime Resistance to the CTX-M β-lactamase (P167S and D240G). However, these mutations actually undergo negative epistasis with one another. Here we show that while P167S and D240G individually grant CTX-M ceftazidime resistance, the double mutant P167S/D240G displays a wild-type behavior. The results presented here suggest that conformational heterogeneity, particularly in active-site loops, plays an important role in the substrate specificity and evolutionary capacity of β-lactamases. The P167S and D240G mutants readily sample multiple conformations and yet are more stable than P167S/D240G, which samples fewer conformations. Crystallography and molecular dynamics results show that the CTX-M acyl-enzyme complex exists in equilibrium between inactive and active conformations and that the P167S and D240G variants have a higher probability of adopting active conformations. Taken together, our data suggests that the P167S and D240G substitutions promote an open conformation of the Ω-loop that creates access for ceftazidime and allows Glu166 to sample conformations consistent with deacylation, whereas the WT and P167S/D240G mutant exhibit a closed Ω-loop conformation that constrains access for ceftazidime and prevents Glu166 from efficiently coordinating water for deacylation.

8.2 Future directions

Altogether, this thesis highlights the importance of understanding allostery in biological systems. Both in scale and approach, we are entering a new phase of using molecular dynamics simulations to understand biophysics; the complex task of simulating an organism's entire proteome could become more commonplace. It is worth speculating what the future holds for

studies of protein dynamics, allosteric communication, and the role of conformational entropy in biological function and disease.

Methodologically, CARDS is already able to measure allosteric communication in a holistic manner, but there are many new methodological enhancements through which we can obtain an increased understanding of allosteric communication. As previously discussed, using a rotameric library might improve upon identifying significant allosteric communication beyond noise, and using a different a multi-exponential distribution to compute probabilities of ordered and disordered regimes might improve detection of dynamical states.

Beyond previously discussed avenues of improvements, construction of our allosteric network in chapter 3 is aided by the knowledge that the GPCR-binding site sends a signal to the GDP-binding site [168–170]. While ample evidence exists to indicate that coupling occurs in both directions, CARDS does not explicitly describe directionality in pairwise communication. This kind of directional information has immense value in less well-studied systems. Hence, there is value in identifying an alternative metric to mutual information that specifically measures a directional transfer of information. Additionally, CARDS extracts data directly from simulation datasets, rather than an MSM. As such, CARDS is inherently limited by the sampling bias that may exist in a dataset, measuring stronger correlations between residues in states that are closer to a trajectory's starting configuration. While computing structural-correlations is already possible using MSMs, it is difficult to assign dynamical states to MSMs. Therefore it is valuable to develop frameworks to extract kinetic signatures of disorder from MSMs that can be used to assign dynamical states. Then correlated motions will be able to be extracted from MSMs trivially.

Along with developments with CARDS, there are a number of questions surrounding allostery in G protein signalling that remain unexplored. While we highlight the potential utility of our allosteric network in chapter 3, it is worth noting that an allosteric inhibitor of Gαq already exists, YM-254890 (YM) [203]. This depsipeptide binds at the hinge between the two domains

and prevents G protein activation by trapping the protein in a GDP bound state, preventing GDP dissociation. However, the mechanism of inhibition, efficacious even in cells [492], remains unclear. Structural models suggest that the domain-opening motion of G proteins is inhibited by YM, but simulation of large peptide-protein complexes, particularly at the scale of heterotrimeric G proteins, remains a challenge. Learning the allosteric mechanism of G protein inhibition by YM might inform chemical strategies to design a simplified peptide analog, as well as inform design principles for the general design of highly-selective G protein inhibitors.

Building off the allosteric networks described in chapters 3 and 4, it is worth exploring if protein homologs have conserved allosteric networks. For example, many G protein isoforms have similar structures [171], but different behaviors ranging from their effector targets to their GDP dissociation kinetics. Given previous success studying different isoforms of a protein family [10], it may be that the differences in conformational landscapes encode similarities and differences in allosteric networks. Assuming a "core" allosteric network exists across all G protein isoforms, the differences between them may be all the more interesting as they might provide a basis for regions that govern isoform-specific behaviors.

Going beyond primarily-simulation studies, it is exciting to enter an era where microsecond-to-millisecond dynamics can be explored and integrated with experiments. While simulations already provide predictive models for experimental measurements (see chapter 3, 4, 5 and 7), there are opportunities to further integrate with experiments that directly measure conformational dynamics, such as single-molecule studies or Nuclear Magnetic Resonance (NMR) spectroscopy. NMR is able to observe major and minor conformational states and obtain thermodynamic and kinetic information at residue-level resolution. It provides a powerful methodology to identify and characterize excited states in a protein's conformational landscape. However, one limitation of NMR is that it remains difficult to characterize these excited states or to identify the motions that may drive the equilibrium motions (called "exchange processes") observed in NMR spectra. MD provides a powerful complement to these limitations, with the potential to characterize the motions underlying conformational equilibria and the participating residues.

NMR and MD together may be a powerful means by which protein conformational landscapes can be more completely characterized. As we enter an era of exascale computing, it is exciting to consider the ways in these kinds of integrated biophysical approaches may help comprehend complex biological phenomena.

Appendix A

Supplementary Material on the CARDS method

The work in this Appendix is published in: Singh, S., and Bowman, G.R., Quantifying allosteric communication via both concerted structural changes and conformational disorder with CARDS. Journal of Chemical Theory and Computation. 13:1507-1517, 2017. PMID: 28282132. [493]

A.1 Supplementary Methods

A.1.1 Molecular dynamics simulations

All simulations were carried out on GROMACS (version 5.1.1) [145, 222] using periodic boundary conditions in a dodecahedron with explicit water solvent. Simulations were carried out at 300K using the AMBER03 [146] force field with the TIP3P water model [143]. The starting conformations of wild-type apo and cAMP-bound CAP were generated by placing crystallographic structures (PDB ID: 4N9H and 1CGP respectively) [129, 130] into separate

dodecahedron boxes that extended 1.0 nm from the protein surface in any direction. Starting conformations fore the S62F variant were generated using the PyMol [144] mutagenesis tool. Each system was then minimized independently with the steepest-descent algorithm until the maximum force fell below 1000 kJ/mol/min using a step size of 0.01nm and a cutoff distance of 1.2nm for the neighbor list, Coulomb interactions, and van der Waals interactions. For equilibration runs, all bonds were constrained with the LINCS algorithm [226] and virtual sites [228] were used to allow for a 4fs time step. As before, cut-offs of 1.0 nm were used for the neighbor list, Coulomb interactions, and van der Waals interactions. The Verlet cutoff scheme was used for the neighbor list, and Particle Mesh Ewald [494] was employed for the electrostatics (with a grid spacing of 0.12nm, PME order 4, and tolerance 1e-5). The stochastic velocity rescaling (v-scale) thermostat [224] was used to hold the temperature at 300K, and the Berendsen barostat [495] was used to bring the system to 1 bar pressure. For the production runs, the position restraint was removed and the Parrinello-Rahman barostat [227] was employed. Conformations were stored every 10 ps. For each system, three 500ns runs were conducted totaling to 1.5μs of aggregate simulation time per system.

Dihedral angles were extracted using the MDTraj Library [148] (v. 1.7). Mappings were all drawn in PyMol (version 1.7) [144], and all figures were constructed in Inkscape (v. 0.48).

A.1.2 Sensitivity analysis

We varied all the cutoff values employed in the CARDS algorithm to ensure the robustness of our results. Specifically, we varied the core width from 60° to 90°, the likelihood ratio cutoff from 1.5 to 5.0, and the neighbor distance cutoff from 2-6Å. Fig. A.1 demonstrates that the communication to the CBD does not change dramatically as these parameters are varied.

A.2 Supplementary Figures

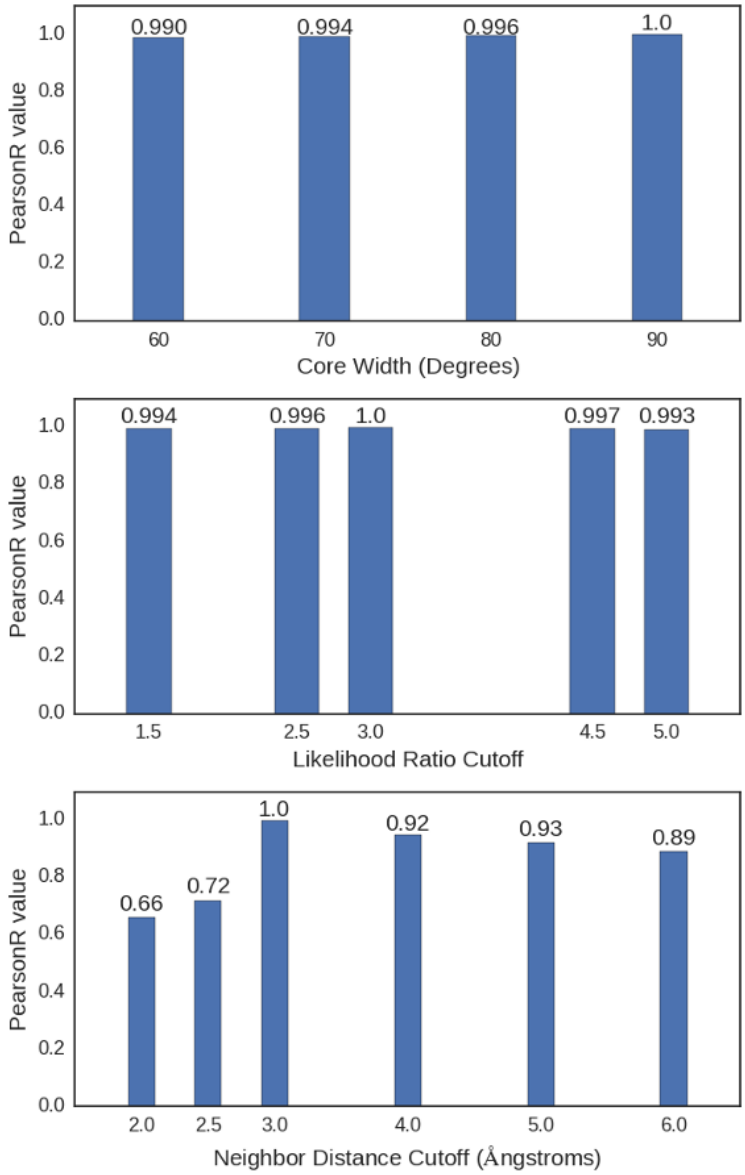


Figure A.1: Pearson Correlation Coefficients (PearsonR) between the CARDS results presented in the main text and those with varying A. the core width, B. the likelihood ratio cutoff, and C. the neighbor distance cutoff.

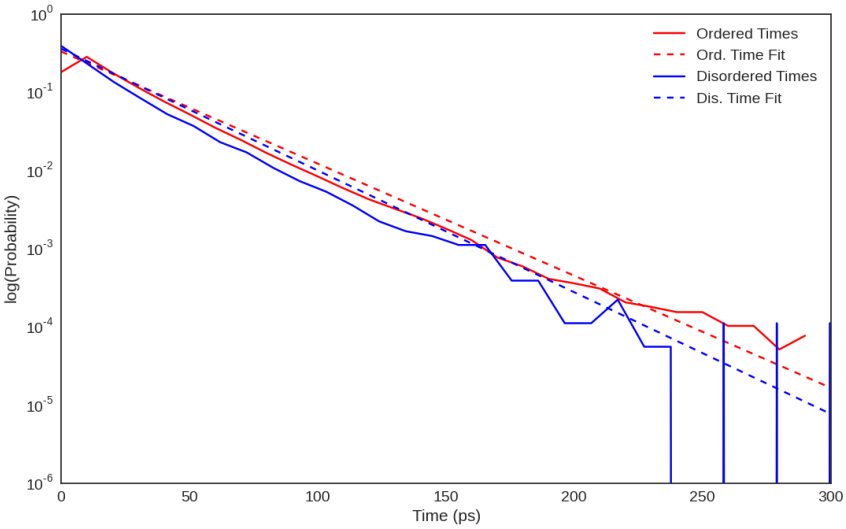


Figure A.2: The distribution of ordered and disordered times for a single dihedral across a single simulation trajectory. The solid lines represent the histogram of times extracted from the trajectory, while the dashed lines represent fits based on the average ordered ($\langle \tau_{ord} \rangle$) and disordered times ($\langle \tau_{dis} \rangle$).

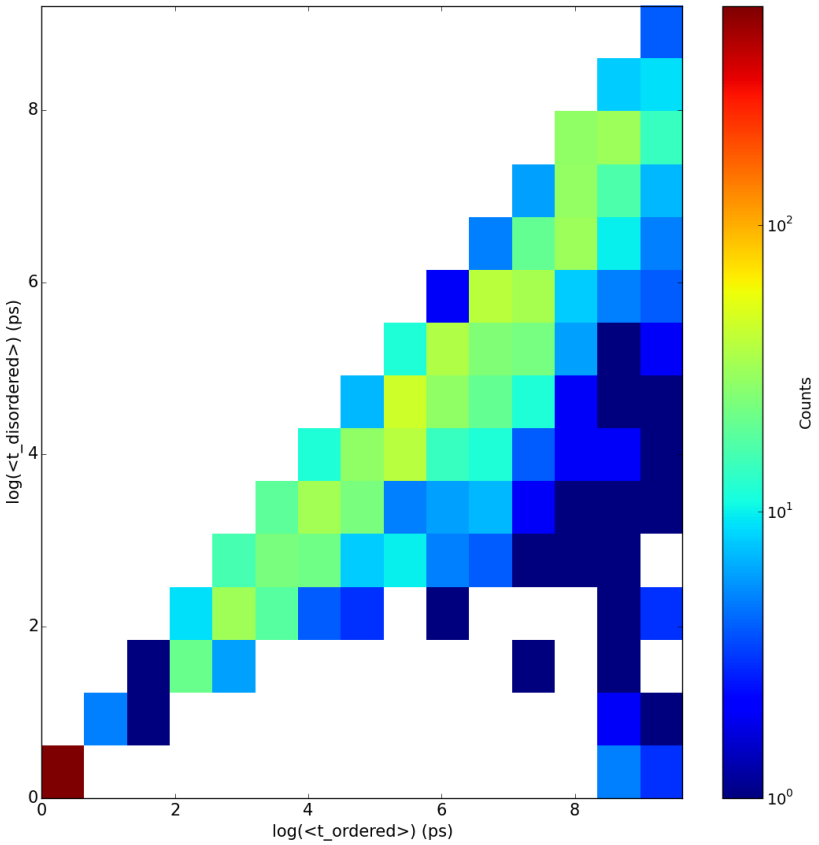


Figure A.3: Two-dimensional histogram of the average ordered and disordered times ($\langle \tau_{ord} \rangle$ and $\langle \tau_{dis} \rangle$) for all dihedrals in CAP.

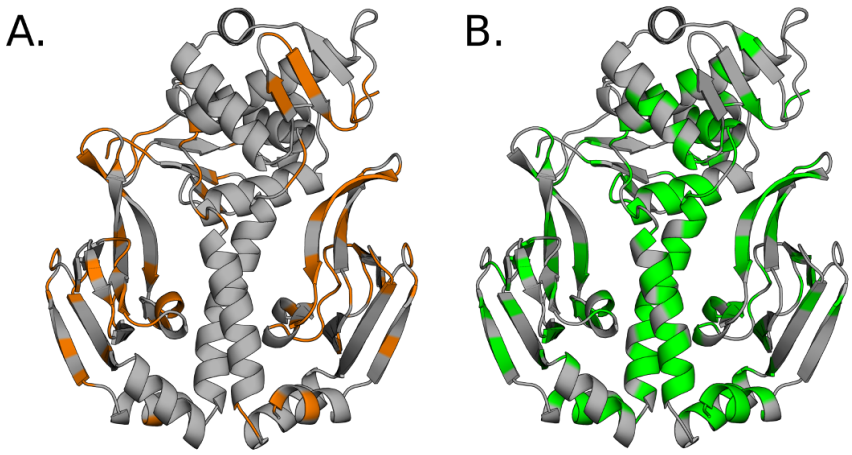


Figure A.4: Residues with separable dihedrals into disordered regimes using a stricter threshold ($\langle\tau_{ord}\rangle \geq 5 \times \langle\tau_{dis}\rangle$) A. Residues with at least one backbone dihedral that is capable of disorder-mediated communication (orange). B. Residues with at least one side-chain dihedral that is capable of disorder-mediated communication (green).

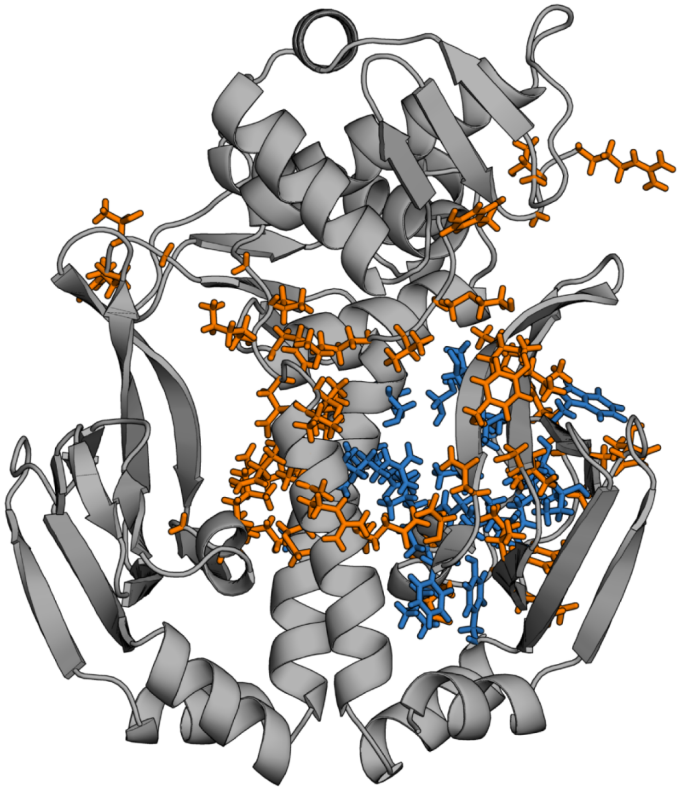


Figure A.5: The top 5% of residues (orange sticks) with disorder-mediated communication to the cAMP-binding pocket (blue sticks). Note that having disorder-mediated communication to the CBD does not preclude the possibility of also having structural communication.

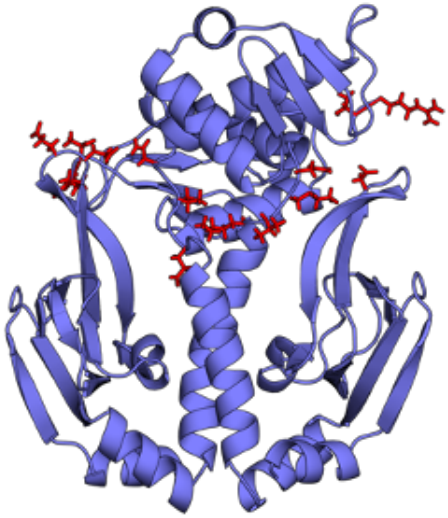


Figure A.6: The top 2% of backbone-side-chain hubs (sticks).

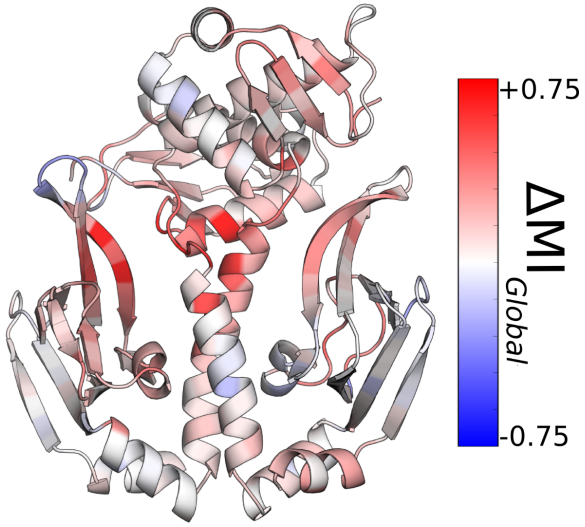


Figure A.7: Change in global communication upon the S62F mutation. The color scale on the right shows the proportional change in global communication relative to the scale in Figure 2.7 of the main text.

Appendix B

Supplementary figures highlighting the mechanism of GDP release

The work in this Appendix is published in: SSun, X. and Singh, S.*, Blumer, K.J., and Bowman, G.R., Simulation of spontaneous G protein activation reveals a new intermediate driving GDP unbinding. eLife, 7, October 2018, https://doi.org/10.7554/eLife.38465.001* [48]

B.1 Supplementary Figures

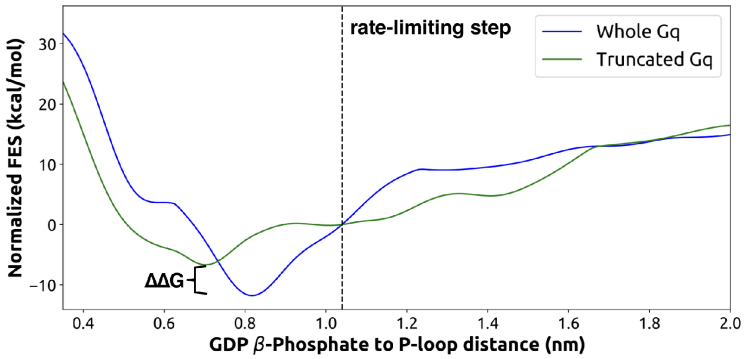


Figure B.1: Free-energy surface from metadynamics simulations of GDP release for the full Gαq (blue) and truncated form (green, without the last five C-terminal residues). Both sets of simulations were run for the same amount of time with identical collective variables. The rate-limiting step as identified from the highest flux pathway is marked with a dashed line. The free-energy difference between the GDP-bound states of the two systems is marked with a bracket.

Table B.1: Measurements comparing tilting and translation of H5 across PDB structures and MD simulation.

Construct description	PDB ID	H5 Tilting Distance (Å)	H5 Vertical translation distance (Å)	H5 Tilting residues used	H5 Transl residues u
Gαq-GDP	3AH8	13.5	10.6	Tyr325 to Leu349	Thr334 to
Gαq after rate limiting step from MD	N/A	15.1	11.1	Tyr325 to Leu349	Thr334 to
Gαi-GDP	1GP2	10.3	10.2	Tyr320 to Ile343	Thr329 to
Gαi-μOR	6DDF	14.6	13.0	Tyr320 to Ile343	Thr329 to
Gαi-A1AR	6D9H	13.8	10.1	Tyr321 to Ile344	Thr330 to
Gαi-Rhodopsin	6CMO	15.8	10.7	Tyr320 to Ile343	Thr329 to
Gαo-5HT1B	6G79	13.1	14.2	Tyr310 to Ile333	Thr319 to
Gαs-B2AR	3SN6	12.8	14.6	Tyr360 to Ile383	Thr369 to

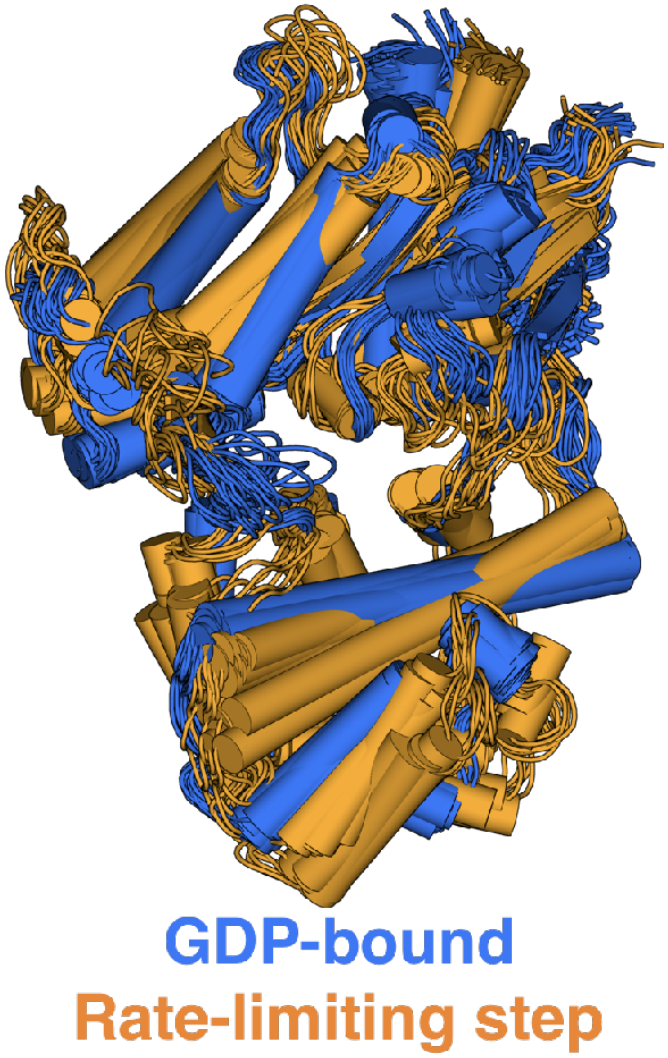


Figure B.2: Overlay of representative structures of Gαq when bound to GDP (blue) or across the rate-limiting step (orange).

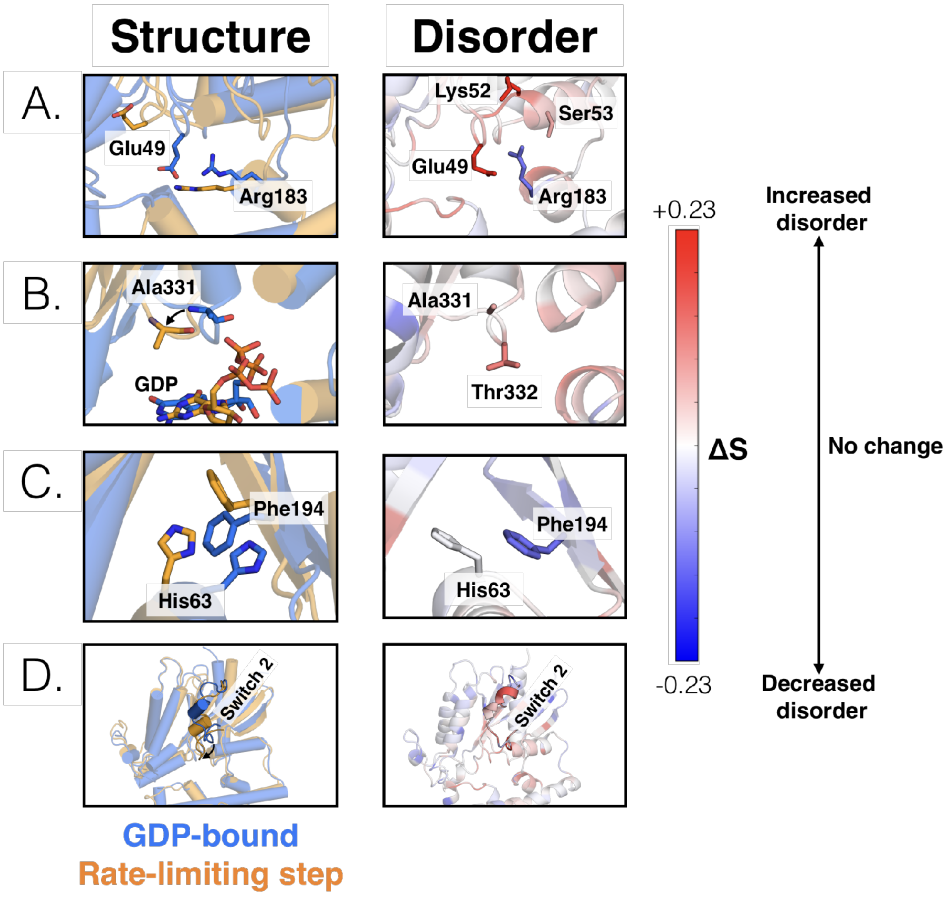


Figure B.3: Changes in the structure (left) and disorder (right) of specific regions across the rate-limiting step. **(A)** Residues that contact the phosphates of GDP, including the salt bridge between Glu49$^{G.s1h1.4}$ and Arg183$^{G.hfs2.2}$, **(B)** the s6h5 loop, **(C)** the $\pi - \pi$ stacking interaction between Phe194$^{G.S2.6}$ and His63$^{G.H1.12}$, and **(D)** switch 2.

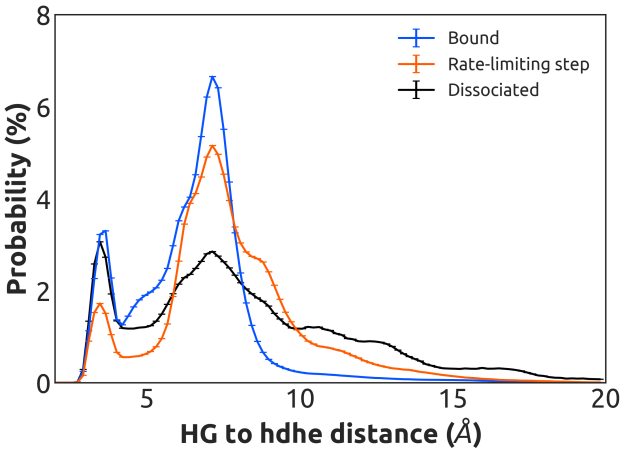


Figure B.4: Distribution of distances between the side-chains of K275$^{G.s5hg.1}$ and D155$^{H.hdhe.5}$ for the GDP-bound state (blue), across the rate-limiting step (orange), and upon GDP dissociation (black).

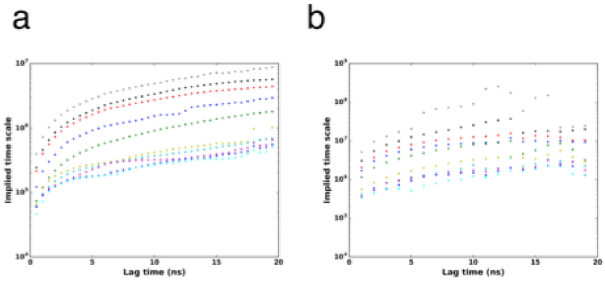


Figure B.5: Implied timescales for the Markov state model. **(A)** Top 10 implied timescales for the 5040 states of Gα. (B) Top 10 implied timescales for the final 221965 states.

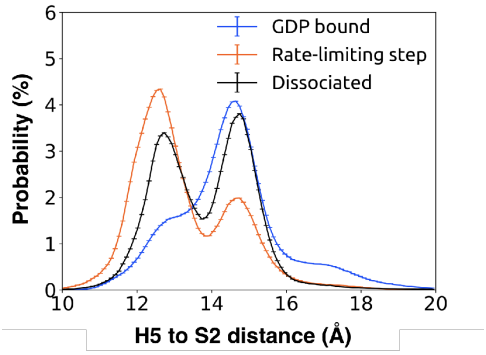


Figure B.6: Probability distribution of the distance between Leu349$^{G.H5.16}$ on H5 and Phe194$^{G.S2.6}$ on S2 to monitor the tilting motion of H5 upon GDP release when bound to GDP (blue), across the rate-limiting step (orange), and upon GDP dissociation (black). In the GDP bound state (blue), such a distance is peaked at 15 Å. Across the rate-limiting step (orange), tilting motion of H5 upon GDP release occurs with a peak in distance at 12.5 Å.

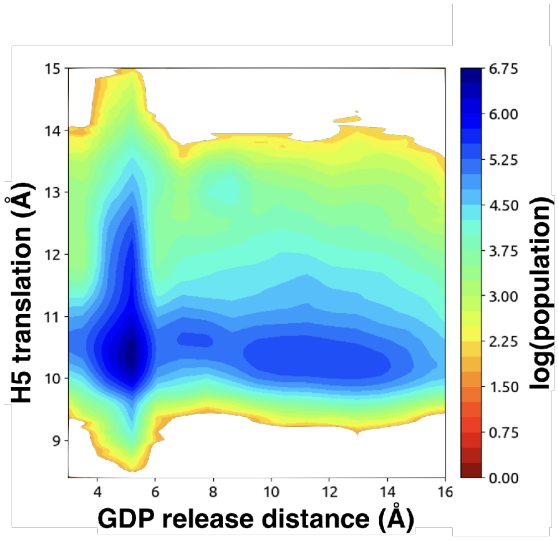


Figure B.7: H5 vertical motion is sampled across GDP release simulations. At each point, the combined population (represented by the color scale) of that state is shown using both GDP-bound and intermediate stages of the GDP release pathway. H5 vertical motion was measured by computing the distance between Thr334$^{G.H5.1}$ on the s6h5 loop and Phe341$^{G.H5.8}$ on H5. GDP release distance was measured as the distance from GDP β-phosphate to the center of mass between residues Lys52$^{G.H1.1}$, Ser53$^{G.H1.2}$, and Thr54$^{G.H1.3}$ on H1.

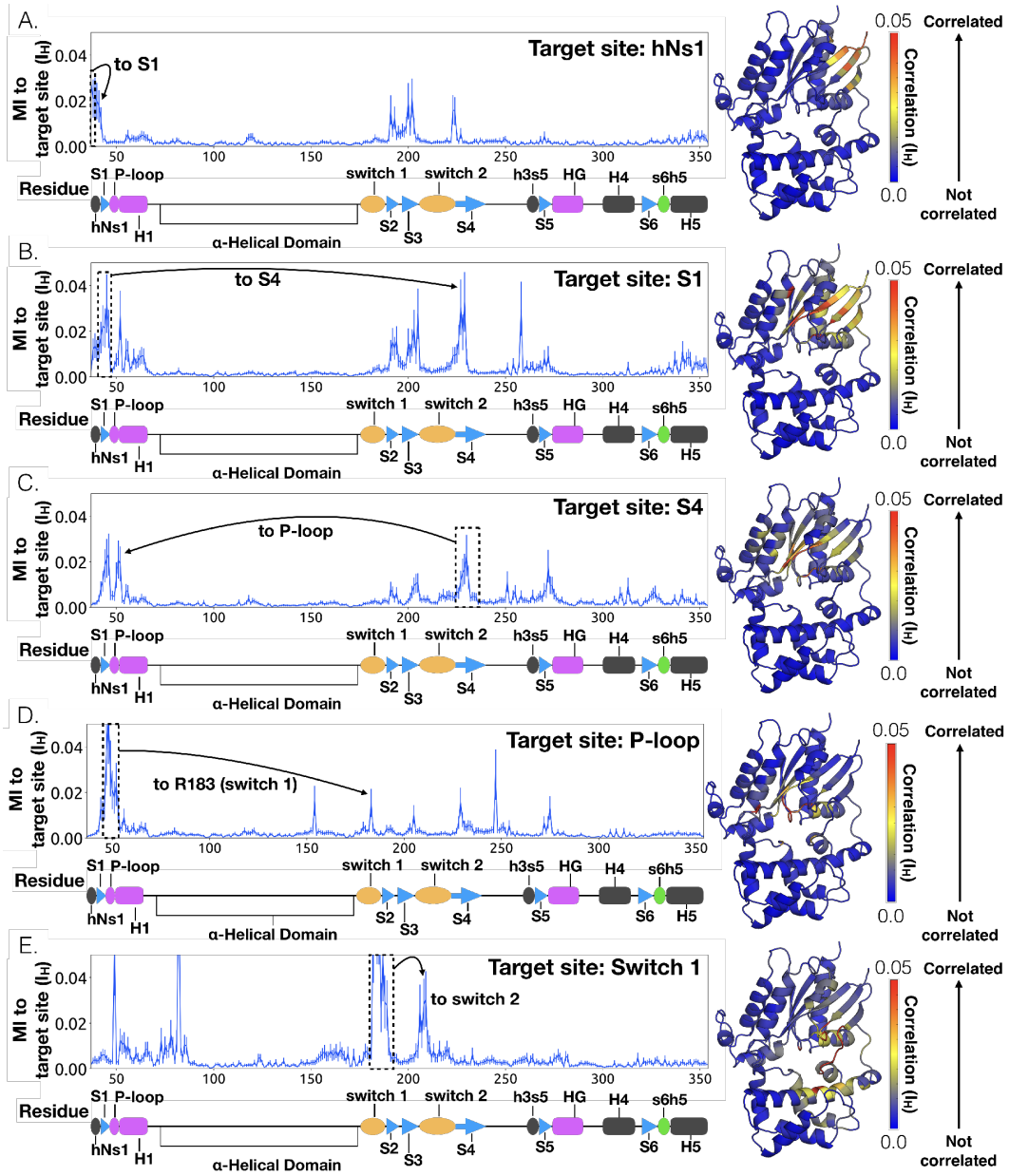


Figure B.8: Allosteric network connecting hNs1 contacts to the P-loop and switch 1 via S4. CARDS data showing communication per residue to a target site (dashed box) is plotted (left) and mapped onto the structure of Gαq (right) for **(A)** hNs1, **(B)** S1 **(C)** S4 **(D)** the P-loop and **(E)** Switch 1. Arrows indicate important regions with significant communication to the target site.

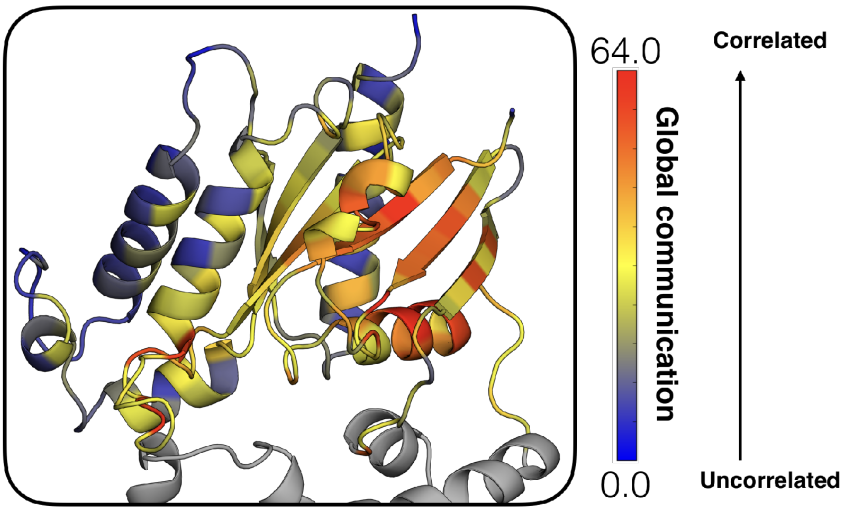


Figure B.9: Global communication of each residue in the Ras-like domain mapped onto the structure of Gαq, colored based on the scale (right). The helical domain (gray) is shown for orientation.

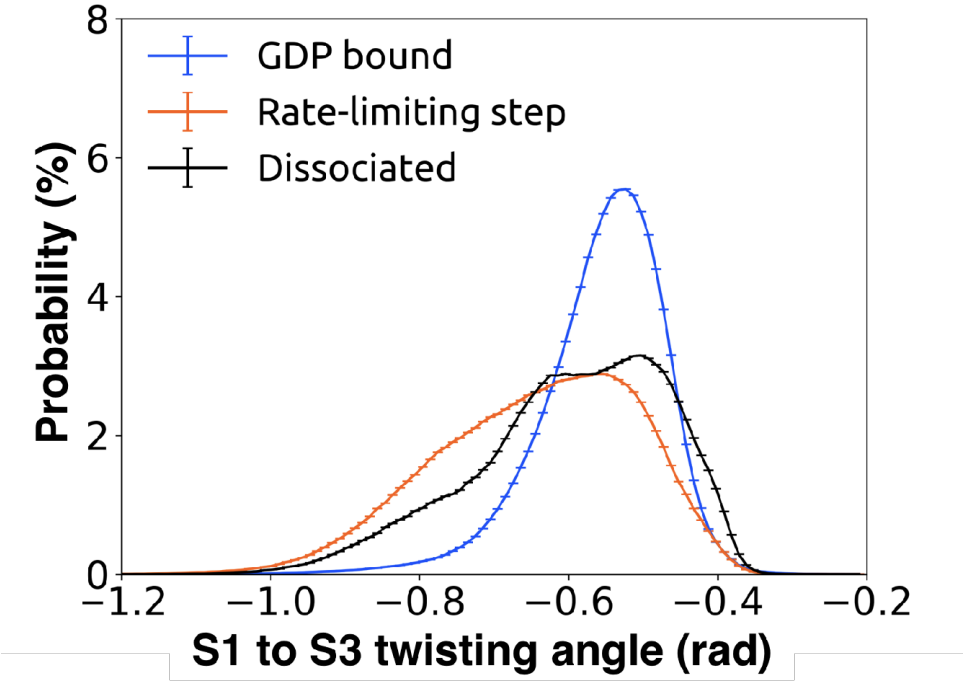


Figure B.10: Probability distributions of the twist angle between S1 and S3. The dihedral angle is computed by taking the dihedral angle between the CA atoms of Leu45$^{G.S1.7}$, Leu40$^{G.S1.2}$, Val199$^{G.S3.1}$, and Asp205$^{G.S3.7}$, so that the angle measured represents S1/S3 twisting at the GPCR facing side. Twist was computed for GDP bound(blue), intermediate(orange), and GDP dissociated states(black).

Appendix C

Supplementary Material to ”Discovery of a cryptic allosteric site in Ebola’s ’undruggable’ VP35 protein”

This chapter is adapted from the following publication:

Cruz, M.A. and Frederick, T.E.*, Singh, S., Vithani, N., Zimmerman, M.I., Porter, J.R., Moeder, K.E., Amarasinghe, G.K., and Bowman, G.R., Discovery of a cryptic allosteric site in Ebola’s ’undruggable’ VP35 protein using simulations and experiments. Preprint on BioRxiv https://doi.org/10.1101/2*
[54]

C.1 Supplementary Material

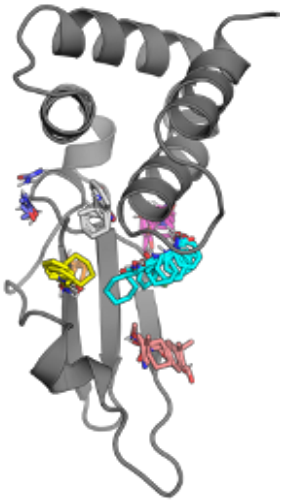


Figure C.1: FTMap results for the main cryptic pocket highlighting an example protein structure (gray) and hotspots where a variety of small organic probes (multicolored sticks) form energetically favorable interactions. The probe molecules are intended to capture different drug-like interactions (such as hydrogen bonding and Van der Waals contacts) and include acetamide, acetonitrile, acetone, acetaldehyde, methylamine, benzaldehyde, benzene, isobutanol, cyclohexane, N,N-dimethylformamide, dimethyl ether, ethanol, ethane, phenol, isopropanol, or urea.

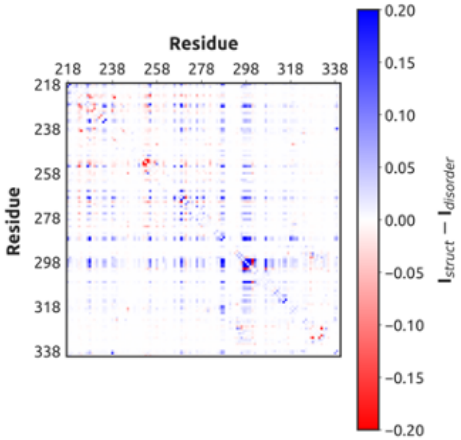


Figure C.2: Purely structural correlations (I_struct) dominate the allosteric network identified by CARDS as they are typically greater than the disorder-mediated couplings ($I_disorder$, which includes correlations between the structure of one residue and the disorder of a second, as well as correlations between the disorder of two residues).

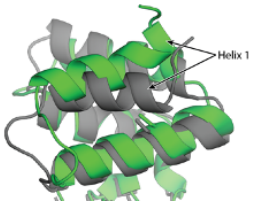


Figure C.3: Motion of helix 1 (green vs gray structures) sometimes exposes C247 (sticks) to solvent. However, the resulting pocket is small and FTMAP does not identify any hotspots in this region that are likely to bind drug-like molecules. Therefore, we focus our attention on the cryptic pocket created by the displacement of helix 7.

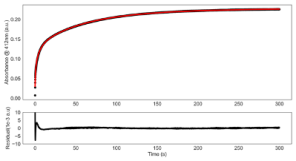


Figure C.4: A representative time trace from a thiol labeling experiment (black) performed at 100 μM DTNB and a quadruple exponential fit (red). The data are background subtracted (e.g. the average absorbance from three runs with DTNB but no protein were subtracted) to account for spontaneous hydrolysis of DTNB.

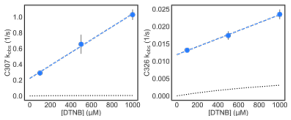


Figure C.5: Thiol labeling of a C247S/C275S variant that only has cysteines in the main cryptic pocket (C307, left and C326, right). Observed labeling rates (blue circles) are shown at a range of DTNB concentrations. Fits to the Linderstrøm-Lang model are shown in dashed colored lines and the expected labeling rate from the unfolded state is shown as black dotted lines. The mean and standard deviation from three replicates are shown but error bars are generally smaller than the symbols.

Variant	C247 rate (s^{-1})	C275S rate (s^{-1})	C307 rate (s^{-1})	C326 rate (s^{-1})
Wild-type	0.011 ± 0.000030	1.4 ± 0.021	0.201 ± 0.0016	0.024 ± 0.00036
C275S	0.0031	-	0.16	0.0061
C275S/C307S	0.012 ± 0.000071	-	-	0.032 ± 0.00017
C275S/C326S	0.012 ± 0.00024	-	0.80 ± 0.0058	-
C247S/C275S	-	-	0.29 ± 0.011	0.013 ± 0.00045
C275S/C307S/C326S	0.025 ± 0.00066	-	-	-
C247S/C275S/C307S	-	-	-	0.0087 ± 0.00027

Figure C.6: Observed labeling rates at 100 μM DTNB for a set of variants with different cysteines mutated to serines to uncover which rate in the wild-type fit corresponds to which cysteine residue. Error is standard deviation from three replicates. Dash represents rates not measured due to the absence of that cysteine residue.

Table C.1: Characterization of the folding/unfolding of VP35's IID used to test whether the observed thiol labeling is due to fluctuations within the native state or global unfolding of the protein. K is the equilibrium constant between the folded and unfolded state determined from denaturation data, kunfold is the unfolding rate of the respective variants measured by intrinsic tryptophan fluorescence.

Variant	K	$K_{unfold}(s^{-1})$
Wild-type	$6.57\times10^{-5} \pm 4.0\times10^{-5}$	0.0175
C247S/C275S	$4.01\times10^{-4} \pm 0.8\times10^{-4}$	0.0083

Table C.2: Intrinsic labeling rates (k_{int}) for each cysteine residue. Intrinsic labeling rates were measured using either urea unfolded variants containing only the specified cysteine, or peptides containing the specified cysteine and its surrounding residues.

Residue	kint$\mu M^{-1}s^{-1}$
C247	0.0566 ± 0.0007
C275	0.00254 ± 0.001
C307	0.0290 ± 0.002
C326	0.395 ± 0.02

Length	Sense Strand	Antisense Strand
25mer	56FAM-rArArArCrUrGrArArArGrGrGrArGrArArGrUrGrArArArGrUrG	rCrArCrUrUrUrCrArCrUrUrCrUrCrCrCrUrUrUrCrArGrUrUrU
25mer with 2nt 3' overhang	56-FAM-rArArArCrUrGrArArArGrGrGrArGrArArGrUrGrArArArGrUrGrCrU	rCrArCrUrUrUrCrArCrUrUrCrUrCrCrCrUrUrUrCrArGrUrUrUrCrU

Figure C.7: RNA sequences used in fluorescence polarization binding assays. The sense and antisense strands were annealed in a 1:1 molar ratio

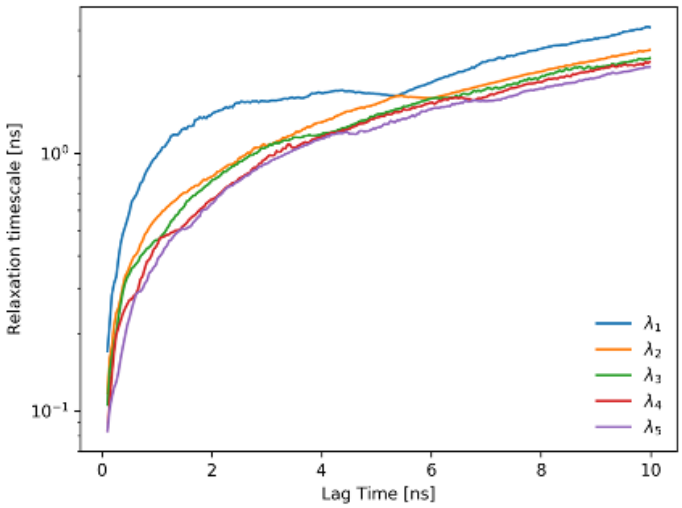


Figure C.8: Implied timescales test for the VP35 IID MSM suggests the kinetics are stable from 3-6ns. Analysis in the main text uses a Markov time of 6 ns. Key results were consistent for lag times from 3-6 ns.

Appendix D

Supplementary Material on the SARS-CoV-2 nucleocapsid protein

This chapter is adapted from the following publication:

Cubuk, J., Alston, J.J., Incicco, J.J., Singh, S., Stuchell-Brereton, M.D., Ward, M.D., Zimmerman, M.I., Vithani, N., Griffith, D., Wagoner, J.A., Bowman, G.R., Hall, K.B., Soranno, A., The SARS-CoV-2 nucleocapsid protein is dynamic, disordered, and phase separates with RNA, Available on Biorxiv: https://doi.org/10.1101/2020.06.17.158121 [2]

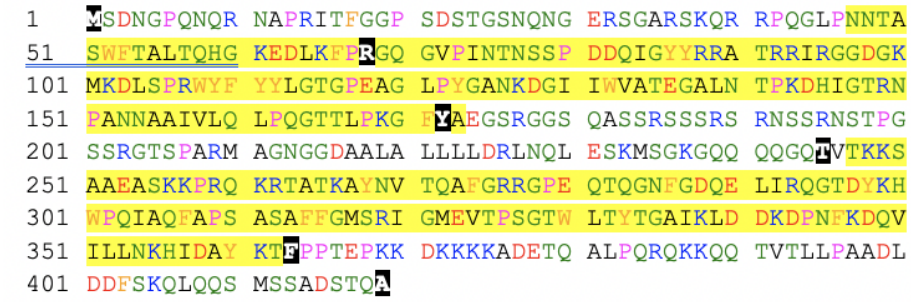
D.1 Supplementary Methods

D.1.1 Sequence analysis

Disorder prediction was performed using IUPred, with additional analysis and sequence parsing done with localCIDER and protfasta, respectively [496–498]

Amino acid sequence of the N protein used in simulations. Highlighted regions delineate folded

domains. Underline bolded residues identify the sites of dyes for single-molecule fluorescence experiments.



D.1.2 Simulation Methods

Monte Carlo simulations

All simulations were performed at 330 K and at 15 mM NaCl, as have been used previously in a variety of systems [338, 401, 496, 499, 499]. Simulation analysis was performed with MDTraj and camparitraj (http://ctraj.com/) [**?**, 148]. For IDR only simulations all degrees of freedom were fully sampled (backbone and sidechain dihedral angles and rigid-body positions) as is standard in CAMPARI Monte Carlo simulations [398]. For simulations of IDRs in the context of folded domains, the backbone dihedral angles of the folded domains were held fixed while all sidechains were fully sampled, as were backbone dihedral angles for the disordered regions, as applied previously [500]. The folded state starting structures were obtained from PDB structures (listed below).

For IDR-only simulations 30-40 independent simulations were run generating final ensembles of 40-60 K conformations. For simulations of IDRs in the context of folded domains, the number of independent simulations and the length of the simulation varied. For the NTD-RBD simulations 400 independent simulations were run, with 2 independent simulations per starting seed from MD simulations (see methods below) leading to a final ensemble of ∼400

K conformations (24 M steps per simulation). For the RBD-LINK-dimerization construct, ten independent simulations were run for a final ensemble of 32 K conformers (66 M steps per simulation). For the dimerization-CTD construct 40 independent simulations were run providing a final ensemble of 40 K conformations (66 M steps per simulation). For a complete description of simulation details see Table D.4.

For the NTD-RB construct, we used a sequential sampling approach in which long timescale MD simulations of the RBD in isolation performed on the Folding@home distributed computing platform were first used to generate hundreds of starting conformations [37, 403]. Those RBD conformations were then used as starting structures for independent all-atom Monte Carlo simulations. Monte Carlo simulations were performed with the ABSINTH forcefield in which the RBD backbone dihedral angles are held fixed but the NTD is fully sampled, as are RBD sidechains. The RBD starting structure used was extracted from the 6VYO PDB crystal structure, which is equivalent to the 6YI3 NMR structure.

For RBD-Link-dimerization domain simulations, we opted to use a single starting seed structure for the folded domains based on the NMR and crystal-structure conformations for the RBD and dimerization domains, respectively. To generate the monomeric starting structure of the dimerization domain, we first built a homology model of the SARS-CoV-2 dimerization dimer from the NMR structure of the SARS dimerization structure (PDB: 2jw8) using SWISS-MODEL [1,306]. We chose this strategy because at the time, no dimerization structure existed, a situation that has since resolved itself [314]. Nevertheless, the SARS and SARS-CoV-2 dimerization domains are essentially identical, such that this is a minor detail.

For dimerization domain-CTD simulations, a single starting structure for the dimerization domain was again used, selected after MD simulations. Having generated a homology model, we extracted a single protomer from the dimeric structure and ran molecular dynamics simulations to identify equilibrated starting structures. In running initial simulations we discovered that as a monomer, the first 21 residues appear disordered, in agreement with sequence predictions

(Fig. D.1A) but in contrast to their behavior in the dimeric structure (Fig. D.1C). As a result, we choose to also model these residues as fully disordered. A single starting seed conformation was used for all dimerization-CTD simulations.

Excluded volume (EV) simulations were performed using the same setup, but with a modified Hamiltonian under which solvation, attractive Lennard-Jones, and polar (charge) interactions are scaled to zero, as described previously [329].

Molecular dynamics simulations

All molecular dynamics simulations of SARS-CoV-2 nucleoprotein were performed with Gromacs 2019 using the AMBER03 force field with explicit TIP3P solvent [143, 146, 501]. Simulations were prepared by placing the starting structure (PDB ID: 6VYO) in a dodecahedron box that extends 1.0 Å beyond the protein in any dimension. The system was then solvated (29125 atoms), and energy minimized with a steepest descents algorithm until the maximum force fell below 100 kJ/mol/nm using a step size of 0.01 nm and a cutoff distance of 1.2 nm for the neighbor list, Coulomb interactions, and van der Waals interactions. For production runs, all bonds were constrained with the LINCS algorithm and virtual sites were used to allow a 4 fs time step [226, 228]. Cutoffs of 1.1 nm were used for the neighbor list with 0.9 for Coulomb and van der Waals interactions. The Verlet cutoff scheme was used for the neighbor list. The stochastic velocity rescaling (v-rescale) thermostat was used to hold the temperature at 300 K [224]. Conformations were stored every 20 ps.

The FAST algorithm was used to enhance conformational sampling and quickly explore the dominant motions of nucleoprotein [35, 42]. FAST-pocket simulations were run for 6 rounds, with 10 simulations per round, where each simulation was 40 ns in length (2.4 μs aggregate simulation). The FAST-pocket ranking function favored restarting simulations from states with large pocket openings. Additionally, a similarity penalty was added to the ranking to promote conformational diversity in starting structures, as has been described previously [278]. The

FAST dataset was clustered using a k-centers algorithm based on RMSD between frames using backbone heavy atoms (C, Cα, Cβ, N, O) to generate 1421 discrete states, which were then launched on the distributed computing platform Folding@home [37, 403].

Furthering conformational sampling and enhancing statistics, Folding@home produced 500 μs of aggregate simulation. A final k-centers clustering was performed with the combined Folding@home and FAST data using Enspara (https://github.com/bowman-lab/enspara) [39]. This clustering was performed the same as described above and generated 200 discrete states that capture maximal diversity in nucleoproteins' conformational ensemble. These states were then used as the basis for CAMPARI simulations.

Sequential MD/MC sampling approach

The NTD and RBD combined are 173 residues of folded and disordered protein, which raises a significant challenge for all-atom sampling. To address this we leveraged a novel approach in which we first ran several microsecond of all-atom molecular dynamics simulations of RBD alone using the Folding@Home platform and the FAST approach for enhanced conformational sampling [35, 37, 403]. We then identified 200 conformationally distinct states based on these simulations which we used as "seeds" for the RBD. Using these seeds, we reconstructed the previously missing NTD and ran all-atom Monte Carlo simulations in which the NTD was fully sampled, the RBD sidechains are fully sampled, but the RBD backbone dihedral angles are held fixed. Multiple replicas of each starting conformation were run, giving us a total ensemble of $\sim$400 K conformations. In parallel, we also ran simulations of the NTD in isolation, enabling an assessment of the impact of the folded domain.

Coarse-grained Polymer Simulations

Coarse-grained simulations were performed using the PIMMS software package [338, 402]. PIMMS is a Monte Carlo lattice-based simulation engine in which each bead engages in anisotropic interactions with every adjacent lattice site. Moves used here were cluster translation/rotation moves and single-bead perturbation moves. Specifically, every simulation step, each bead in the system is sampled to move to adjacent sites in random order 503 of times multiplied by a factor that reflects the length of the chain. Every 100 moves (on average) a cluster of chains is randomly selected and translated or rotated, where a cluster reflects a collection of two or more chains in direct contact. This moveset provides changes to the system that reflect physical movements expected in a dynamical system, allowing us to - for equivalently sized systems - compare the apparent dynamics of assembly, as has been done previously [502–505]. We repeated the simulations presented using a range of different movesets and, while convergence varied from set-to-set, we always observed analogous results.

All simulations were performed in a 70 x 70 x 70 lattice-site box using period boundary conditions. The results reported are averaged over the final 20% of the simulation to give average values after equivalent numbers of MC steps. The "polymer" is represented as a 61-residue polymer with either a central high-affinity binding site or not. The binder is a 2-bead species. Every simulation was run for 20 x 10^9 Monte Carlo steps, with four independent replicas. Simulations were run with 1,2,3,4 or 5 polymers and 50, 75, 100, 125, 150, 175, 200, 250, 300, 400 binders.

If our simulations are run in a way deliberately designed to rapidly reach equilibrium using enhanced sampling approaches eventually all single-polymer condensates coalesce into one large multi-polymer condensate. Hence, our simulations are deliberately designed explore a regime in which single-polymer condensates are metastable.

D.1.3 Protein expression, purification, and labeling.

Plasmid construct design.

SARS-CoV2 Nucleocapsid protein (NCBI Reference Sequence: YP_009724397.2) including an N term extension containing His9-HRV 3C protease site – CATCATCACCATCATCATCAT-CACCACCTCGAAGTTCTGTTCCAAGGCCCGATGAGTGATAACGGTCCCCAGAATCAACG GAATGCGCCCAGAATCACGTTCGGCGGTCCAAGCGACAGTACAGGTTCGAATCAGAATGGTGAAC AGCAAACAGCGTCGTCCACAGGGTTTGCCGAACAATACGGCTAGCTGGTTCACTGCGCTGACGCA ACTTAAAATTTCCGCGAGGCCAGGGGGTCCCGATTAATACTAACTCCTCCCCTGACGATCAAATTG TGCAACCCGCCGTATCCGCGGCGGAGACGGTAAAATGAAAGATCTGTCACCGCGCTGGTATTTTTA GGTCCTGAAGCAGGCTTGCCGTATGGCGCTAACAAAGATGGCATTATCTGGGTGGCTACCGAGGGT CGAAAGATCATATTGGAACCCGTAACCCAGCCAATAACGCAGCAATCGTACTGCAGCTGCCGCAGC GAAAGGCTTTTATGCGGAAGGGAGTCGTGGCGGCAGCCAAGCCAGCTCCCGTAGCTCCTCGCGCT CGGAATAGTACACCGGGTTCATCACGCGGCACCTCGCCGGCACGCATGGCTGGCAACGGGGGGGA TACTTTTACTGGATAGGCTTAACCAGTTGGAAAGTAAAATGAGCGGTAAAGGCCAGCAGCAGCAGG CAAAAAGAGCGCGGCAGAGGCGTCGAAAAAACCTAGACAAAAGCGTACTGCGACCAAAGCCTAC TTCGGCCGGCGCGGTCCGGAACAAACCCAGGGCAACTTTGGTGACCAGGAGCTGATTCGTCAGGG ACTGGCCACAGATCGCGCAATTTGCCCCCTCGGCGTCAGCCTTTTTTGGTATGTCTCGCATTGGGAT GTCTGGCACGTGGCTGACGTACACGGGCGCTATAAAGCTGGATGATAAAGATCCGAACTTCAAAGA CTGAACAAACATATTGACGCCTATAAAACGTTCCCCCCTACTGAACCTAAGAAAGATAAAAAAAAA CCCAAGCGCTACCACAACGCCAGAAAAAGCAGCAGACCGTCACCCTCCTGCCGGCAGCGGACCTG GCAACTGCAACAAAGCATGTCAAGCGCCGATAGTACACAGGCGTAA - was cloned into the BamHI EcoRI sites in the MCS of pGEX-6P-1 vector (GE Healthcare) to express the protein product:

GST-LEVLFQGPLGS**HHHHHHHHH**

LEVLFQGPMSDNGPQNQRNAPRITFGGPSDSTGSNQNGERSGARSKQRRPQGLPNNTASWFTALTQH PINTNSSPDDQIGYYRRATRRIRGGDGKMKDLSPRWYFYYLGTGPEAGLPYGANKDGIIWVATEGAL

ANNAAIVLQLPQGTTLPKGFYAEGSRGGSQASSRSSSRSRNSSRNSTPGSSRGTSPARMAGNGGDAAL
ESKMSGKGQQQQGQTVTKKSAAEASKKPRQKRTATKAYNVTQAFGRRGPEQTQGNFGDQELIRQGT
SASAFFGMSRIGMEVTPSGTWLTYTGAIKLDDKDPNFKDQVILLNKHIDAYKTFPPTEPKKDKKKKA
QQTVTLLPAADLDDFSKQLQQSMSSADSTQA. Site-directed mutagenesis was performed on the His9-SARS-CoV2 Nucleocapsid pGEX vector to create M1C R68C, Y172C T245C, and F363C A419C variant N protein constructs. All cloning and site-directed mutagenesis steps were performed by Genewiz and sequences were verified using sanger sequencing.

Protein Expression and Purification.

Both GST-His9-SARS-CoV2 M1C-R68C and Y172C-T245C Nucleocapsid variants were expressed recombinantly in BL21 Codon-plus pRIL cells (Agilent). 4L cultures were grown in LB medium containing carbenicillin (100 ug/mL) to OD600 $\sim$ 0.6 and induced with 0.2 mM IPTG for 12 hours at 16ºC. Harvested cells were lysed with sonication at 4ºC in lysis buffer (50mM Tris pH 8, 500 mM NaCl, 10% glycerol, 10 mg/mL lysozyme, 5 mM BME, cOmplete™ EDTA-free Protease Inhibitor Cocktail (Roche), DNAse I (NEB), RNAse H (NEB)). The supernatant was cleared by centrifugation (37000 rpm for 1 hr) and bound to an HisTrap FF column (GE Healthcare) in buffer A (50 mM Tris pH 8, 500 mM NaCl, 10% glycerol, 20mM imidazole, 5 mM BME). GST-His9-N protein fusion was eluted with buffer B (buffer A + 500 mM imidazole) and dialyzed into cleavage buffer (50 mM Tris pH 8, 50 mM NaCl, 10% glycerol, 1 mM DTT) with HRV 3C protease, thus cleaving the GST-His9-N fusion yielding FL N protein with two additional N-term residues (GlyPro). FL N protein was then bound to an SP sepharose FF column (GE Healthcare) and eluted using a gradient of 0-100% buffer B (buffer A: 50mM Tris pH 8, 50mM NaCl, 10% glycerol, 5 mM BME, buffer B: buffer A + 1 M NaCl) over 100 min. Purified N protein variants were analyzed using SDS-PAGE and verified by electrospray ionization mass spectrometry (LC-MS). Concentrations were determined spectroscopically in 50mM Tris (pH 8.0), 500mM NaCl, 10% (v/v) glycerol using an extinction coefficient = 42530 $M^{-1}cm^{-1}$

Both GST-His9-SARS-CoV2 wild-type and F363C A419C Nucleocapsid variants were expressed recombinantly in Gold BL21(DE3) cells (Agilent). 4 L cultures were grown in LB medium with carbenicillin (100 ug/mL) to OD600 $\sim$ 0.6 and induced with 0.2 mM IPTG for 12 hours at 16ºC. Harvested cells were lysed with sonication at 4ºC in lysis buffer (listed above). The supernatant was cleared by centrifugation (37000 rpm for 1 hr) and the pellet was resuspended in 50 mM Tris pH 8, 500 mM NaCl, 10% glycerol, 6 M Urea, 5 mM BME and incubated at 4ºC for one hour. The resuspension was cleared by centrifugation (37000 rpm for 1hr) and the GST-His9-N protein in the supernatant was bound to a FF HisTrap column (GE Healthcare) in buffer A (50 mM Tris pH 8, 500 mM NaCl, 10% glycerol, 20 mM imidazole, 5 mM BME) containing 6 M Urea. The column was then washed with buffer A allowing the protein to refold on the column. The GST-His9-N protein fusion was then eluted with buffer B (buffer A containing 500 mM imidazole) and dialyzed into cleavage buffer (50 mM Tris pH8, 50 mM NaCl, 10% glycerol, 1 mM DTT) containing HRV 3C protease. FL N protein was then bound to an SP sepharose FF column (GE Healthcare) and eluted using a gradient of 0-100% buffer B (buffer A: 50mM Tris pH 8, 50 mM NaCl, 10% glycerol, 5 mM BME, buffer B: buffer A + 1 M NaCl) over 100 min. Purified N protein variants were analyzed using SDS-PAGE and verified by electrospray ionization mass spectrometry (LC-MS). Protein concentrations of stock solutions were determined spectroscopically in 50mM Tris (pH 8.0), 500mM NaCl, 10% (v/v) glycerol using an extinction coefficient = 42530 $M^{-1}cm^{-1}$

Fluorescent Dye Labeling.

All Nucleocapsid variants were labeled with Alexa Fluor 488 maleimide (Molecular Probes) under denaturing conditions in buffer A (50mM Tris pH8, 50mM NaCl, 10% glycerol, 6M Urea, 1mM DTT) at a dye/protein molar ratio of 0.7/1 for 2 hrs at room temperature. Single labeled protein was isolated via ion-exchange chromatography (Mono S 5/50 GL, GE Healthcare - protein bound in buffer A and eluted with 0-100% buffer B (buffer A + 1 M NaCl) gradient over 100 min) and UV-Vis spectroscopic analysis to identify fractions with 1:1 dye:protein

labeling. Single labeled Alexa Fluor 488 maleimide labeled N protein was then subsequently labeled with Alexa Fluor 594 maleimide at a dye/protein molar ratio of 1.3/1 for 2 hrs at room temperature. Double labeled (488:594) protein was then further purified via ion-exchange chromatography (Mono S 5/50 GL, GE Healthcare - see above).

D.1.4 Single Molecule Spectroscopy

Experimental setup and procedure.

Single-molecule fluorescence measurements were performed with a Picoquant MT200 instrument (Picoquant, Germany). For single-molecule FRET measurements, a diode laser (LDH-D-C-485, PicoQuant, Germany) was synchronized with a supercontinuum laser (SuperK Extreme, NKT Photonics, Denmark), filtered by a z582/15 band pass filter (Chroma) and pulsed at 20 MHz for pulsed interleaved excitation (PIE) (Müller et al., 2005) of labeled molecules. Emitted photons were collected with a 60x1.2 UPlanSApo Superapochromat water immersion objective (Olympus, Japan), passed through a dichroic mirror (ZT568rpc, Chroma, USA), and filtered by a 100 μm pinhole (Thorlabs, USA). Photons are counted and accumulated by a HydraHarp 400 TCSPC module (Picoquant, Germany). For FRET-FCS measurements, the same diode laser was used in continuous-wave mode to excite the donor dye. Photons emitted from the sample were collected by the objective, and scattered light was suppressed by a filter (HQ500LP, Chroma Technology) before the emitted photons passed the confocal pinhole (100 mm diameter). The emitted photons were then distributed into four channels, first by a polarizing beam splitter and then by a dichroic mirror (585DCXR, Chroma) for each polarization. Donor and acceptor emission was filtered (ET525/50m or HQ642/80m, respectively, Chroma Technology) and then focused on SPAD detectors (Excelitas, USA). The arrival time of every detected photon was recorded with a HydraHarp 400 TCSPC module (PicoQuant, Germany).

FRET experiments were performed by exciting the donor dye with a laser power of 100 μW

(measured at the back aperture of the objective). For pulsed interleaved excitation experiments, the power used for exciting the acceptor dye was adjusted to match a total emission intensity after acceptor excitation to the one observed upon donor excitation (between 50 and 70 mW). Single-molecule FRET efficiency histograms were acquired from samples with protein concentrations between 50 pM and 100 pM. Trigger times for excitation pulses (repetition rate 20 MHz) and photon detection events were stored with 16 ps resolution.

For fluorescence correlation spectroscopy (FCS) experiments, acceptor-donor labeled samples with a concentration of 100 pM were excited by either the 485 nm diode laser or the supercontinuum laser at the powers indicated above. However, in the experiments on protein oligomerization, due to an increase in the fluorescence background upon addition of unlabeled protein above 1 μM, only the correlations corresponding to direct acceptor excitation (582 nm) have been considered reliable for the analysis.

For nsFCS, FRET samples of acceptor-donor labeled protein with a concentration of 100 pM were excited by the same diode laser but in continuum wavelength mode.

All measurements were performed in 50 mM Tris pH 7.32, 143 mM β-mercaptoethanol (for photoprotection), 0.001% Tween 20 (for surface passivation) and GdmCl at the reported concentrations. All measurements were performed in uncoated polymer coverslip cuvettes (Ibidi, Wisconsin, USA), which significantly decrease the fraction of protein adhering to the surface (compared to normal glass cuvettes) under native conditions. For comparison, experiments have been performed also in glass cuvette coated with PEG, which provided analogous results to the polymeric cuvette.

Each sample was measured for at least 30 min at room temperature (295 $\pm$ 0.5 K).

FRET efficiency histograms.

Fluorescence bursts from individual molecules were identified by time-binning photons in bins of 1 ms and retaining the burst if the total number of photons detected after donor excitation was larger than 30. Transfer efficiencies for each burst were calculated according to $E = nA/(nA + nD)$ where nD and nA are the number of donor and and receptor photons, respectively. Corrections for background, acceptor direct excitation, channel crosstalk, differences in detector efficiencies, and quantum yields of the dyes were applied (Schuler et al., 2012). The labeling stoichiometry ratio S was computed accordingly to

$$S = \frac{I_D}{\gamma_{PIE}(I_A + I_D)} \tag{D.1}$$

where $I_D$and I_A represent the total intensities observed after donor and acceptor excitation and γ_{PIE} provides a correction factor to account for differences in the detection efficiency and laser intensities. Bursts with stoichiometry corresponding to 1:1 donor:acceptor labeling (in contrast to donor and acceptor only populations) were selected and finally from the selected bursts a histogram of transfer efficiencies is constructed. Variations in the selection criteria for the stoichiometry ratio do not impact significantly the observed mean transfer efficiency (within experimental errors).

To estimate the mean transfer efficiency and deconvolve multiple populations (e.g for the NTD construct) from the transfer efficiency histograms, each population was approximated with a Gaussian peak function. For fitting more than one peak, the histogram was analyzed with a sum of Gaussian peak functions. For the conversion of transfer efficiency to distances, we used the value of the Förster radius for Alexa488 and Alexa594 previously determined and reported in literature, R_0 = 5.4 nm [506]. We further correct the value accounting for the dependence of the Förster radius on the solution refractive index. The changes in refractive index caused by increasing concentrations of GdmCl or KCl were measured with an Abbe refractometer (Bausch and Lomb, USA).

Finally, we estimated a systematic error on transfer efficiency of $\pm$ 0.03, based on the variation of transfer efficiency of the same reference samples after different calibrations of the instrument over the last two years. Standard deviation of the transfer efficiency for multiple repeats of the NTD, LINK, and CTD constructs is equal or less than $\pm$ 0.01. Since we aim for a comparison with simulations, here we consider the systematic error as the larger source of error and we propagate the corresponding effect on all the calculated distances.

Fluorescence lifetimes and anisotropies analysis.

A quantitative interpretation of this transfer efficiency in terms of distance distribution requires the investigation of protein dynamics. A first method to assess whether the transfer efficiency reports about a rigid distance (e.g. structure formation or persistent interaction with the RBD) or is the result of a dynamic average across multiple conformations is the comparison of transfer efficiency and fluorescence lifetime.The interdependence of these two factors is expected to be linear if the protein conformations are identical on both timescales (nanoseconds as detected by the fluorescence lifetime, milliseconds as computed from the number of photons in each burst). Alternatively, protein dynamics give rise to a departure from the linear relation and an analytical limit can be computed for configurations rearranging much faster than the burst duration. The dependence of the fluorescence lifetimes on transfer efficiencies determined for each burst was compared with the behavior expected for fixed distances and for a chain sampling a broad distribution of distances. For a fixed distance, R, the mean donor lifetime in the presence of acceptor is given by

$$t_D(R) = t_{D0}(1 - E(R)) \tag{D.2}$$

where t_D is the lifetime in the absence of acceptor, and

$$E(R) = \frac{1}{1 + \frac{R^6}{R_0^6}} \tag{D.3}$$

For a chain with a dye-to-dye distance distribution $P(R)$, the donor lifetime is

$$t_D = \frac{\int tI(t)dt}{\int I(t)dt} \tag{D.4}$$

where

$$I(t) = I_0 P(R) \exp\left(\frac{-t}{tD(R)}\right)dR \tag{D.5}$$

is the time-resolved fluorescence emission intensity following donor excitation. A similar calculation can be carried out for describing the acceptor lifetime [507] delay given by

$$\frac{t_A(R) - t_{A0}}{t_{D0}} \tag{D.6}$$

Donor and acceptor lifetimes at different concentrations of GdmCl were analyzed by fitting subpopulation-specific time-correlated photon counting histograms after donor and acceptor excitation, respectively.

Multiparameter detection allows also excluding possible artifacts, such as insufficient rotational averaging of the fluorophores or quenching of the dyes. Subpopulation-specific anisotropies were determined for both donor and acceptor of all three constructs for NTD, LINK, and CTD, and values were found to vary between 0.1 and 0.2 for the donor and between 0.1 and 0.2 for the acceptor, sufficiently low to assume as a good approximation for the orientational factor $\kappa^2 = 2/3$.

Fluorescence Correlation Spectroscopy (FCS) analysis.

In order to determined changes in the hydrodynamic radius (Rh) of the protein, FCS correlations were analyzed assuming 3D diffusion of the molecule across a three dimensional Gaussian profile of the confocal volume (Rigler, Eur Blophys J (1993) 22:169-175). For 1 diffusing species, and in the absence of photophysical transitions in the time scale of the lag times ana-

lyzed, this formalism amounts to the following time autocorrelation function.

$$g(\tau) = t + \frac{1}{N}(1 + \frac{\tau}{\tau_D})^{-1}(1 + \frac{\tau}{\alpha^2\tau_D})^{-1/2} \tag{D.7}$$

where N is the average number of molecules in the confocal volume, τD is the diffusion time along the xy plane, α is the eccentricity of the three dimensional Gaussian observational volume.

$$\tau_D = \frac{\omega_{xy}^2}{4D} \tag{D.8}$$

where D is the 3D translational diffusion coefficient and ωxy is the radius from the center of the laser beam at which the light intensity decreases e^2 times from its maximum value at the center. α = ωz/ωxy.

Additionally, in order to account for contributions of the photophysics of the fluorophore to the correlation observed in the μs timescale, we added two triplet terms multiplying the diffusion correlation term (see for example Krichevsky, Rep. Prog. Phys. 65 (2002) 251–297). The overall equation that we fit to the FCS traces is then

$$g(\tau) = 1 + (g_D(\tau) - 1)(1 + c_{T1}\exp(-\frac{\tau}{\tau_{T1}}))(1 + c_{T2}\exp(-\frac{\tau}{\tau_{T2}})) \tag{D.9}$$

where τ_{T1}, $\tau_{T2}, c_{T1}, and c_{T1}$, denotes the characteristic times and amplitudes of the contributions of two triplet states to$g(\tau)$. Parameters $\tau_D, \tau_{T1}, \tau_{T2}, c_{T1}, c_{T2} and N$ were fitted by least square nonlinear regression analysis for each concentration of unlabeled protein tested (Fig. D.13A-B), while α was fixed at a value of 6 determined independently from analysis of fluorescence intensity profiles of fluorescent nanobeads.

Making use of the definition of τ_D and the Stokes-Einstein equation, we have, for each concentration of unlabeled protein

$$\frac{\tau_D}{\tau_{D0}} = \frac{R_h}{R_{h0}} \tag{D.10}$$

where τ_{D0} and R_{h0} are the diffusion time and hydrodynamic radius in the absence of unlabeled

protein, respectively. Error bars in Fig. D.13 B are the standard errors of Rh / Rh0 estimated from propagation of the standard errors across multiple measurements of the diffusion times obtained from the fit.

Nanosecond Fluorescence Correlation Spectroscopy.

Autocorrelation curves of acceptor and donor channels and cross-correlation curves between acceptor and donor channels were calculated with the methods described previously [330,508]. All samples have been measured at a concentration of 100 pM and bursts with a transfer efficiency between 0.3 and 0.8 have been selected to eliminate the contribution of donor only to the correlation amplitude. Finally, the correlation was computed over a time window of 5 μs and characteristics timescales were extracted according to:

$$g_{ij} = 1 + \frac{1}{N}(1 - c_{AB}\exp[-\frac{\tau - \tau_0}{\tau_{AB}}])(1 + c_{CD}\exp[-\frac{\tau - \tau_0}{\tau_{CD}})(1 + c_T\exp[-\frac{\tau - \tau_0}{\tau_T}) \quad \text{(D.11)}$$

where N is the mean number of molecules in the confocal volume and i and j indicate the type of signal (either from the Acceptor or Donor channels). The three multiplicative terms describe the contribution to amplitude and timescale of photon antibunching (AB), chain dynamics (CD), and triplet blinking of the dyes (T). τ_{CD} is then converted in the reconfiguration time of the interdye distance τ_r correcting for the filtering effect of FRET as described previously [509]. An additional multiplicative CD term has been added only for the donor-donor correlations to describe the fast decay observed at very short time. Such a decay is not found in the correlations of other disordered proteins measured on the instrument and we associate the fast decay with the rotational motion of the overall protein. A fit to this fast decay is about 2 ns.

Polymer models of distance distributions.

Conversion of mean transfer efficiencies for fast rearranging ensembles requires the assumption of a distribution of distances. Here, we compared the results of two distinct polymer models: the Gaussian model and a Self-Avoiding Walk (SAW) model that accounts for changes in the excluded volume [510]. This second model has been shown to provide a better description of chain distribution and scaling exponent when compared to distance distributions from MD simulations [511]. Importantly, both models rely only on one single fitting parameter, the root mean square interdye distance $r = \langle R^2 \rangle^{1/2}$ for the Gaussian chain and the scaling exponent ν for the SAW model.

Estimates of these parameters are obtained by numerically solving:

$$\langle E \rangle = \int_0^{l_c} P(R)E(R)dr \tag{D.12}$$

where R is the interdye distance, l_c is the contour length of the chain, $P(r)$ represents the chosen distribution, and $E(R)$is the Förster equation for the dependence of transfer efficiency on distance R and Förster radius:

$$E(R) = \frac{R_0^6}{R_0^6 + R^6} \tag{D.13}$$

The Gaussian chain distribution is given by:

$$P_{FJC}(R, r) = 4\pi R^2 (\frac{3}{2\pi r^2})^{3/2} \exp(-\frac{3R^2}{2r^2}) \tag{D.14}$$

The SAW model can be expressed as:

$$P_{SAW}(R, \nu) = A_1 \frac{4\pi}{b_0 N^\nu} (\frac{R}{b_0 N^\nu})^{2+g} \exp(-A_2 (\frac{R}{b_0 N^\nu})^\delta) \tag{D.15}$$

where

$$A_1 = \frac{\delta}{4\pi}\frac{\Gamma[5+\frac{g}{\delta}^{\frac{3+g}{2}}]}{\Gamma[3+\frac{g}{\delta}^{\frac{5+g}{2}}]}, A_2 = (\frac{\Gamma[5+\frac{g}{\delta}]}{\Gamma[5+\frac{g}{\delta}]})^{\delta}, g = \frac{\gamma-1}{\nu}, \delta = \frac{1}{1-\nu} \tag{D.16}$$

γ =1.1615, and Γ is the Euler Gamma Function, b_0=0.55nm is an empirical prefactor [511], N is the number of residues between the fluorophores, and ν is the scaling exponent.

Finally, when converting the distance from transfer efficiencies, to account for the length of dye linkers and compare the experimental data with simulations, the root-mean-squared interdye distance r was rescaled according to

$$r_{m,n} = |m-n|^{0.5} I_{dye} |m-n+2I_{dye}|^{0.5} \tag{D.17}$$

with I_dye=4.5 (Aznauryan et al., 2016; Hoffmann et al., 2007). Finally, the persistence length is computed using the Gaussian conversion $r^2 = 2l_p l_c$ [348].

Binding of denaturant and folding.

As in previous works [318, 322, 512], we model the chain expansion with the denaturant in terms of a simple binding model:

$$r_c = r_0(1+\rho\frac{Kc}{1+Kc}) \tag{D.18}$$

Where r_0 is the mean square interdye distance at zero denaturant,ρ is a term the captures the extent of chain expansion with the denaturant compared to r_0, and the K is the binding constant, and c is the concentration of denaturant.

In presence of folded domains, we can imagine the folding/unfolding of the domains can affect the overall size of the chain because of an increase or decrease of excluded volume due to the surrounding folded domains (which screen part of the available conformations) or because of the folding or unfolding of elements in the region between the fluorophores. To account for

this effect, as in the case of the NTD, we weighed the effect of denaturant on the chain for the fraction folded f_f and unfolded f_u accordingly to:

$$r_c = (r_0 f_f + r_{0u} f_u)(1 + \rho \frac{Kc}{1 + Kc}) \tag{D.19}$$

where r_{0f} and r_{0u} are the root mean square interdye distance in presence of folded or unfolded domains in native buffer,

$$f_f = \frac{\exp[-m(c - c_m)]}{1 + \exp[-m(c - c_m)]} \tag{D.20}$$

and $f_u = 1 - f_f$, where c_m the midpoint concentration and m the denaturant m value, representing the dependence of free energy on denaturant concentration. The stability parameter ΔG_0 can be computed as $\Delta G_0 = m c_m$.

Polymer model of electrostatic interactions.

The disordered regions of the N protein are enriched in positive and negative charges. To provide a term of comparison in the interpretation of protein conformations as function of salt concentration, we use the polymer theory for polyampholyte solutions developed by Higgs and Joanny [512, 513], which has been shown previously to capture quantitatively the conformational changes of unstructured proteins. Briefly, the root mean square interdye distance is equal to $r = N^{0.5} * l_0 * \alpha$ where N is the number of monomers in the disordered region, l_0 is the length of elementary segment (here 0.36 nm) and α is the ratio between l and l_0, with l being a rescaled segment that accounts for excluded volume and electrostatic interactions.

$\alpha$is computed according to the equation proposed by Higgs and Joanny [512, 513]:

$$\alpha^5 - \alpha^3 = \frac{4}{3}(\frac{3}{2\pi})^{1.5} N^{0.5} v^* \tag{D.21}$$

where v^* is an effective excluded volume given by the sum of three terms:

$$v^* b^3 = v b^3 + \frac{4\pi l_B (f-g)^2}{k^2} - \frac{\pi l_B^2 (f-g)^2}{k} \tag{D.22}$$

Here, v is the excluded volume (accounting for physical excluded volume and positive and attractive interactions that are not due to electrostatics), f and g are the fraction of positive and negative residue respectively for considered segment of the protein, k is the Debye screening length, and l_b is the Bjerrum length.

Importantly, when accounting for the fraction of negative charges, we also account for the contribution of the -2 net charge of each dye at pH 7.3.

Salt dependence of NTD, LINK, and CTD conformations.

In addition to studying the conformations under native buffer conditions, we investigate how salt affects the conformations of the three disordered regions. We started by testing the effects of electrostatic interactions on the NTD conformational ensemble. Moving from buffer conditions and increasing concentration of KCl, we observed a small but noticeable shift toward lower transfer efficiencies, which represents an expansion of the NTD due to screening of electrostatic interactions. This can be rationalized in terms of the polyampholyte theory of Higgs and Joanny [512, 513] (see Table D.2), where the increasing concentration of ions screens the interaction between oppositely charged residues (see Fig. D.10).

We then analyzed for comparison the LINK construct. Interestingly, we find a negligible effect of salt screening on the root mean square distance r172-245 as measured by FRET (see Fig. D.10). Predictions of the Higgs and Joanny theory for the content of negative and positive charges within the LINK construct indicates a variation of interdye distance dimension that is comparable with the measurement error. It has to be noted that in this case the excluded volume term in the Higgs and Joanny theory will empirically account not only for the excluded volume

of the amino acids in the chain, but also for the excluded volume occupied by the two folded domains.

Finally, we test if the addition of salt can provide similar effects than those obtained by GdmCl on the conformations of the CTD: interestingly, we do not observe any significant variation either in transfer efficiency or distribution width (Fig. D.10), suggesting that the broadening of the population observed for the CTD does not originate from electrostatic interactions.

D.1.5 Testing protein oligomerization.

NativePAGE experiments were performed to verify that purified recombinantly expressed SARS-CoV-2 N protein is capable of forming dimers and oligomers, in analogy to SARS-CoV N protein, and as shown in more recent work for SARS-CoV-2 [303, 306, 316]. Indeed, NativePAGE experiments reveal the existence of multiple bands (Fig. D.13C-D). However, since the lowest band in the NativePAGE corresponds to an apparent molecular weight of $\sim$70-80 kDa, we wanted to verify the oligomeric state of this band.

To test whether the apparent mass is due to a slow mobility of the protein because of its high positive charge, we performed crosslinking experiments. These experiments confirm the formation of dimers, tetramers, and high oligomeric species, as a function of protein concentration above 500 nM (Fig. D.13E-F). These oligomeric species are in equilibrium with the monomer, the smallest species on the denaturing SDS PAGE (which has the expected molecular weight of $\sim$45 kDa). It has to be noted that, because of the slow reactivity of the crosslinking agent (see Methods below), the crosslinking experiments do not represent the population of monomeric and oligomeric species at equilibrium. However, the comparison between the NativePAGE and the crosslinking experiments supports the fact that the smallest band in the NativePAGE is indeed the monomer protein.

We finally turned to Fluorescence Correlation Spectroscopy (FCS) to test whether labeled pro-

tein can form dimers. We measured the CTD construct that carries one labeling position at the end of the oligomerization domain. When increasing the concentration of unlabeled protein, we observe a systematic increase in the hydrodynamic radius when compared to the hydrodynamic radius under native conditions (Fig. D.13A-B). This suggests that the labeled protein can form higher oligomeric species in a concentration regime comparable to the one observed in NativePAGE and SDS PAGE experiments and that at 100 pM (the concentration used in single-molecule experiments), no oligomer is formed. Caution must be used in the interpretation of the oligomeric bound species observed in FCS experiments, since labeling mutation may have affected the affinity of the dimerization domain. Future experiments will address the role of mutation on dimerization. Finally, all experiments have been performed at two different time points, after 1 hour and after 24 hours of incubation of the labeled sample with unlabeled protein to test any kinetic effect on the measured value. No significative difference has been observed.

Taken together, NativePAGE crosslinking experiments verify that in smFRET and FCS experiments we are in fact monitoring the behavior of the monomeric SARS-CoV-2 N protein.

D.1.6 Protein Crosslinking Methods.

50 mM disuccinimidyl suberate (DSS) (Thermo Scientific) stock solution was prepared (10 mg into 540 uL of anhydrous DMSO (Sigma)). All protein samples were prepared in 20 mM NaPi pH 7.4 (with and without 200 mM NaCl) at the following concentrations: 0.1, 0.5, 1, 5, 10 and 20uM. DSS stock solution was added to each sample to a final concentration of 1.25 mM. Samples were incubated for 1 hour at room temperature. Samples were then quenched to a final concentration of 200mM Tris pH 7.4 and allowed to incubate for 15 minutes. Crosslinked proteins were then analyzed using SDS PAGE and Coomassie staining.

D.1.7 NativePAGE Methods.

All protein samples were prepared in 20 mM NaPi pH 7.4 (with and without 200 mM NaCl) at the following concentrations: 0.05, 0.1, 0.5, 1, 5, 10 and 20 uM. Samples were subjected to NativePAGE (Invitrogen) and protein mobility was analyzed with Coomassie staining.

Development of turbidity in solutions of N protein and poly(rU) was followed through measurements of absorbance at 340 nm in a microvolume spectrophotometer (NanoDrop, Thermo, USA). Mixtures were prepared in 500 μl plastic reaction tubes by adding 4 μl protein solution into 3 μl of poly(rU) and absorbance was recorded 45 s – 75 s after mixing. Working solutions were kept at room temperature during experiments.

Reaction media was 50 mM Tris, pH 7.5 (HCl), 0.002 % v/v Tween20, and NaCl as indicated in Results.

poly(rU) (Midland Certified Reagent Company, TX, USA, lot number 011805) was reconstituted into this media from stocks dissolved in RNAse free water. According to the manufacturer, the size of poly(rU) molecules is mostly less than 250 nucleotides (nt.) and longer than 200 nt.

Protein stocks (in 50 mM Tris pH 8.0, 500 mM NaCl, 10% v/v glycerol) were buffer exchanged into the desired buffer through size exclusion chromatography in Zeba Spin 7 k MWCO desalting columns (Thermo, USA). poly(rU) concentrations in working dilutions were assessed through the absorbance at 260 nm employing an extinction coefficient of 9.4 $mM^{-1}cm^{-1}$ (Michelson, 1959). Protein concentrations were assessed through the absorbance at 280 nm employing an extinction coefficient of 42.53 $mM^{-1}cm^{-1}$, computed according to the method proposed by Pace et al. (Pace et al., 1995).

The limiting concentrations of nucleic acid across which an increase in turbidity was detected were estimated through interpolation of the data. To this end, an empirical equation, describing the trends observed at all concentrations, was fitted to the data and then was solved to extract the

poly(rU) concentrations at which turbidity reaches a limit value above the background signal. We used a limiting absorbance value of 0.005 units (340 nm, 1 mm path length). We found that an appropriate function for this end is an exponential of a Gaussian distribution function $F(x)$:

$$F(x) = A(1 - \exp[-\beta\gamma(x)]) \tag{D.23}$$

where

$$\gamma(x) = \frac{1}{(2\pi)^{0.5}\sigma} \exp\,[\frac{-(x-\mu)^2}{2\sigma}] \tag{D.24}$$

where x denotes poly(rU) concentration and $A, \mu, \sigma and \mu$ are parameters fitted through weighted minimum least squares for each protein concentration (solid lines in Fig. 5.5A-B and limiting value points in panels C-D). To characterize the observed global trends of turbidity, as a function of both RNA and protein concentration, we determined approximate functional forms of the dependence on protein concentration of the individually fitted parameters ($A(p), \beta(p), \sigma(p) and \mu(p)$, where p is protein concentration). The observed dependencies were increasing linearly for $\mu(p)$ and quadratic for $\beta(p)$ and $\sigma(p)$. A was the worst defined parameter and thus displayed the least clear trend. For the results in absence of added salt we employed an increasing power function with exponent as a fitting parameter (best fit value was < 1), whereas for the results in presence of 50 mM NaCl the trend of $A(p)$ was better described by a decreasing exponential function.

We thus used the functional forms $A(p), \beta(p), \sigma(p)$ and $\mu(p)$ to construct a global function dependent on both protein and RNA concentration. Global fitting of this equation to the whole set of turbidity titration curves provided the turbidity contour plots shown in Fig. 5.5C-D (solid lines). Contour lines were computed at 1, 10, 20, 50 and 100 times the limiting value employed ($A_{340nm,1mm}$= 0.005).

D.2 Supplementary Figures

Table D.1: Fit parameters to denaturant binding model.

	ρ	K (M^{-1})	r_0 (Å)
NTD (1 pop)	1.3± 0.2	0.36 ± 0.05	50 ± 2 (fixed)
NTD (2 pop)	(shared parameter)	(shared parameter)	36 ± 3
LINK	1.1 ± 0.03	0.06 ±0.03	57 ± 2 (fixed)
CTD	0.47 ± 0.02	0.36 ± 0.1	50 ± 2 (fixed)

Table D.2: Fit parameters of Higgs & Joanny theory

	v
NTD	4.4 ± 0.1
LINK	5.5 ± 0.3
CTD	8.4 ± 0.9

Table D.3: Scaling exponents

	νSAW	νsimulation
NTD	0.520 ± 0.009	0.52
LINK	0.538 ± 0.008	0.58
CTD	0.546 ± 0.008	0.49

Table D.4: All-atom simulation summary

System	No. sims	Total steps per sim (M).	Prod. steps per sim.(M)	Config. output	Ensemble size
NTD-RBD	400	24	20	20,000	399,000
RBD-LINK-DIM	10	66	60	20,000	31,113
DIM-CTD	40	24	20	20,000	40,000
NTD	40	71	66	30,000	64,000
LINK	30	101	80	30,000	66,660
CTD	40	71	66	30,000	64,000

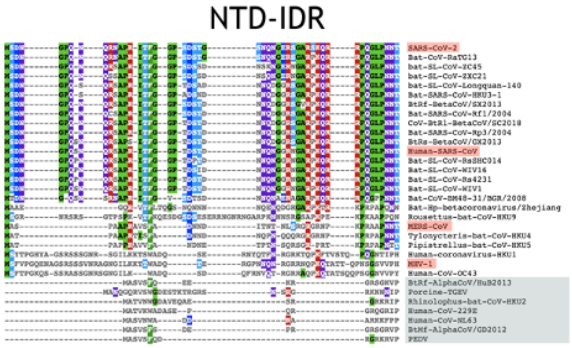


Figure D.1: Sequence alignment of the coronavirus N-terminal domain (NTD).

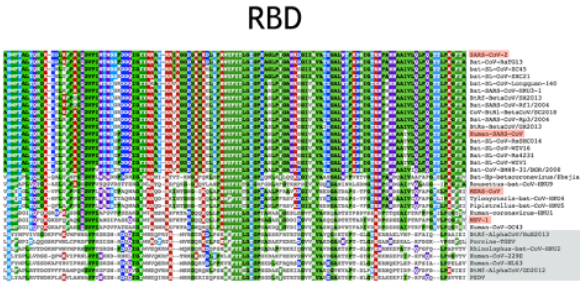


Figure D.2: Sequence alignment of the coronavirus RNA binding domain (RBD).

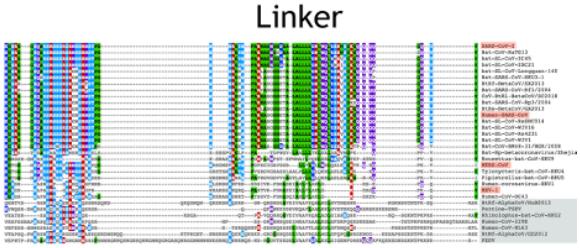


Figure D.3: Sequence alignment of the coronavirus linker (LINK).

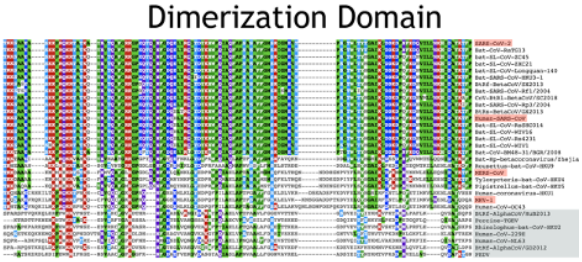


Figure D.4: Sequence alignment of the coronavirus dimerization domain.

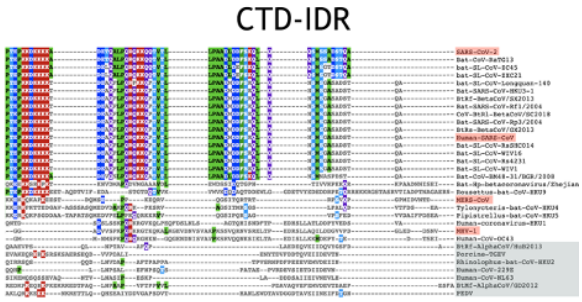


Figure D.5: Sequence alignment of the coronavirus C-terminal domain (CTD)

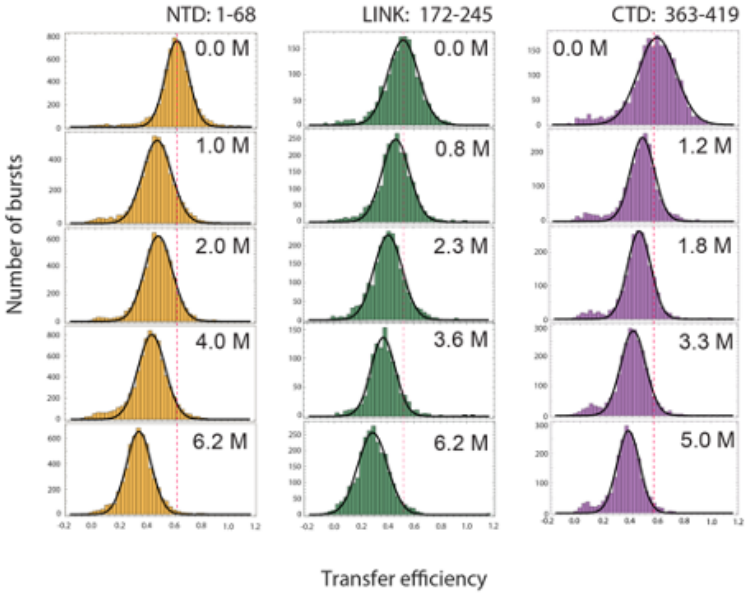


Figure D.6: Histograms of transfer efficiency distributions across denaturant concentrations for NTD, LINK, and CTD constructs.

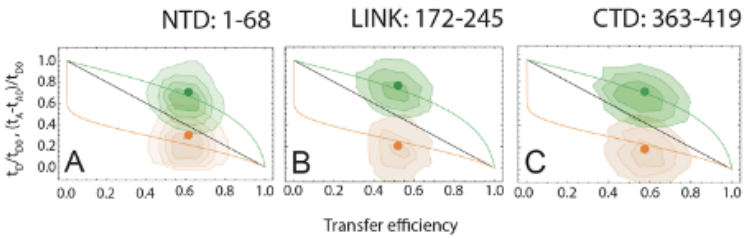


Figure D.7: Dependence of fluorescence lifetime on transfer efficiency. **A.** NTD construct. **B.** LINK construct. **C.** CTD construct. Black line: linear dependence expected for a rigid molecule. Green line: the donor lifetime (normalized by the donor lifetime in absence of FRET: tD/tD0) in the limit of dynamics much faster than the burst duration but slower than the fluorophore lifetime. Orange line: the acceptor lifetime delay (normalized by the donor lifetime in absence of FRET. The green and orange contour plots represent the corresponding distributions of donor lifetime and acceptor lifetime delay as observed in single-molecule experiments under native conditions The green and orange dots represent the mean value of the measured distributions. A larger overlap between donor and acceptor lifetime populations is observed for the NTD and CTD, hinting to possible static conformations.

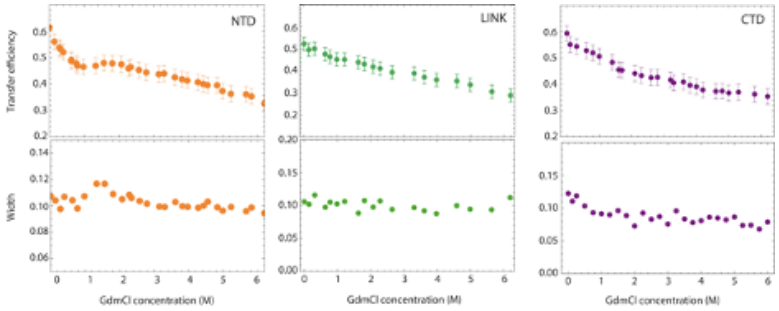


Figure D.8: Mean transfer efficiency and width of NTD, LINK, and CTD across denaturant. The mean transfer efficiency of the NTD domain exhibits a plateau between 1 and 2 M; at the same concentration we observe a small but systematic increase in the amplitude of the transfer efficiency distribution hinting to the coexistence of two populations in slow exchange with very similar transfer efficiencies. The CTD width also shows a small increase in the width of the transfer efficiency distribution that may reflect the formation of local structure under native conditions (e.g. the putative helical binding motif).

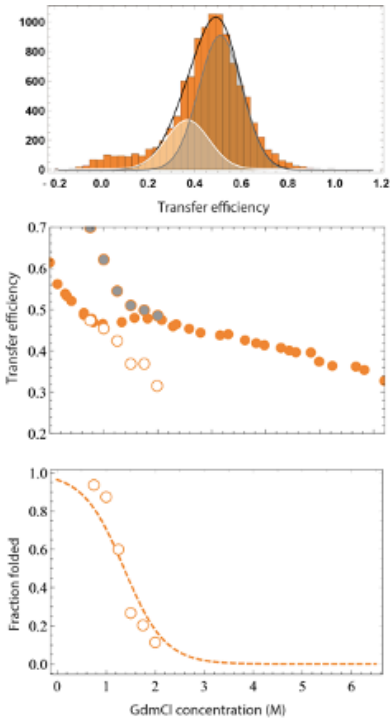


Figure D.9: Fit of NTD construct with two populations. To address the change in amplitude that occurs from the NTD construct between 1 and 2 M GdmCl, we attempt a fit of the same data using two populations with a fixed distance equal to average width outside the 1-2 M GdmCl region (see for comparison Fig. D.8). **Upper panel:** fit of the transfer efficiency histogram at 1.5 M GdmCl. The white- and gray- shaded areas reflect fits to the "folded RBD" population and to the "unfolded RBD" population. **Central panel:** Comparison of transfer efficiencies with a single fit (solid orange circles, compare Fig. D.8) and from the two populations: gray solid circles for the "unfolded RBD" population and unfilled circles for the "folded RBD" population. **Lower panel:** Fraction folded estimated from the fit with Eq. S7 compared to the fraction of "folded RBD" obtained from computing the ratio between the area under "folded RBD" species and the total histogram area.

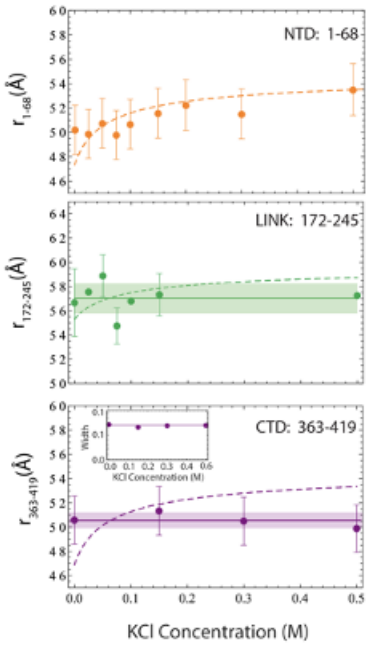


Figure D.10: Interdye distances of NTD, LINK, CTD in presence of salt (KCl). **Upper panel:** root mean square interdye distance between position 1 and 68. Dashed line: fit according to the Higgs and Joanny model (Eq. D.21-D.22) predicts a comparable change to the one observed. Central panel: root mean square interdye distance between position 172 and 245. Dashed line: fit according to the Higgs and Joanny model (Eq. D.21-D.22) predicts a comparable change to the one observed. Solid line and shaded area: average value of the root-mean-square interdye distance across all salt conditions and corresponding standard deviation. The standard deviation is comparable to the measurement error. **Lower panel:** root mean square interdye distance between position 363 and 419. Dashed line: fit according to the Higgs and Joanny model (Eq. D.21-D.22) does not capture the observed trend. This can be possibly explained considering the significant predicted population of helical conformations in the CTD. Solid line and shaded area: average value of the root-mean-square interdye distance across all salt conditions and corresponding standard deviation. Inset: no variation in the width of the transfer efficiency population is observed upon addition of KCl.

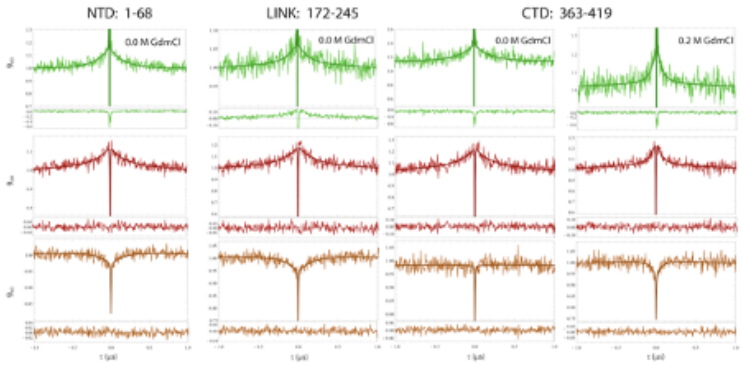


Figure D.11: Chain dynamics measured via ns-FCS. Nanosecond FCS measurements for the NTD, LINK, and CTD constructs provide a measure of the dynamics on the nanosecond timescale. The donor-donor (green), acceptor-acceptor (red), and donor-acceptor (orange) correlation are fitted to a global model that accounts for antibunching, FRET dynamic populations, and triplet. The acceptor-donor correlation shows a clear anticorrelated change for NTD and LINK in the signal that reflects the anticorrelated nature of the donor-acceptor energy transfer as a function of distance: an increase in acceptor reflects a decrease in donor. The CTD cross-correlation exhibits a flat behavior. Occurrence of a correlation in the donor-donor and acceptor-acceptor autocorrelations corresponding with a characteristic time t_b= 190 $\pm$ 30 ns suggests the presence of chain dynamics, either through FRET or Photo-induced Electron Transfer (PET) [331, 332]. Addition of 0.2 M GdmCl, which causes a small decrease in the transfer efficiency width (Fig. D.8) leads to an anticorrelation in the cross-correlation of CTD. The correlation decay appears also faster with a $t_C D$= 70 $\pm$ 15 ns. All measurements are normalized to the value measured at 1 μs for highlighting the amplitude relative to the reconfiguration term.

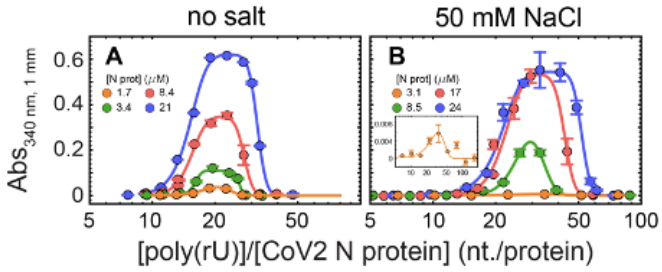


Figure D.12: Turbidity experiments plotted against RNA/protein ratio. Representative turbidity titrations with poly(rU) in 50 mM Tris, pH 7.5 (HCl) at room temperature, in absence of added salt **(A)** and in presence of 50 mM NaCl **(B)**, at the indicated concentrations of N protein. On the x-axis, the concentration of poly(rU) is rescaled for the protein concentration. Points and error bars represent the mean and standard deviation of 2-4 consecutive measurements from the same sample. Solid lines are simulations of an empirical equation fitted individually to each titration curve. An inset is provided for the titration at 3.1 μM N protein in 50 mM NaCl to show the small yet detectable change in turbidity on a different scale. Interestingly, within the experimental error, we observe a clear alignment of the turbidity curves with a maximum at 20 nucleotides per protein in the absence of added salt **(A)** and 30 nucleotides per protein in the presence of 50 mM NaCl **(B)**.

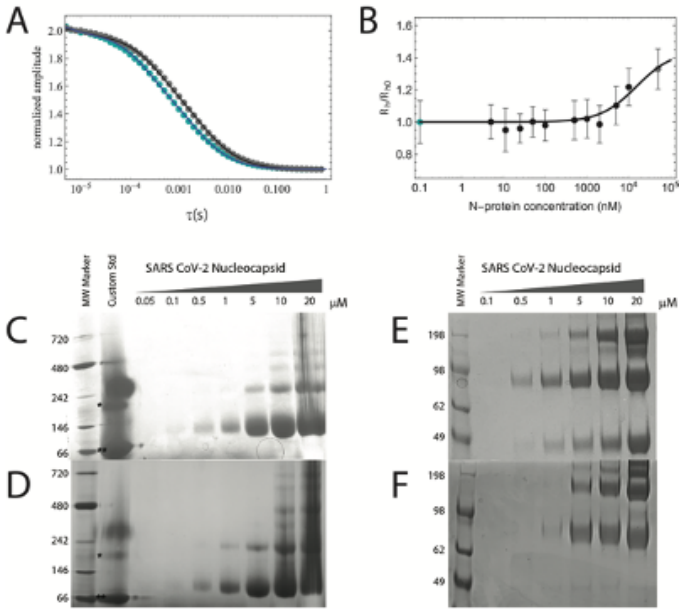


Figure D.13: Testing SARS-CoV-2 N protein oligomerization. **(A-B)** Fluorescence Correlation Spectroscopy (FCS) of full-length SARS-CoV-2 N protein as a function of protein concentration. (A) FCS traces of 100pM Alexa 488/Alexa 594 N protein labeled at positions 363 and 419 in the absence (blue dots) and the presence (gray dots) of 50 μM unlabeled N protein. (B) Hydrodynamic radius of SARS-CoV-2 N protein obtained from FCS trace analysis (blue dot: 100pM labeled N protein; gray dot: 100pM labeled N protein + 50μM unlabeled N protein). **(C-D)** NativePAGE of full-length SARS-CoV-2 N protein in 20 mM NaPi pH 7.4 as a function of protein concentration in the presence of 200 mM NaCl (C) and in the absence of added salt (D). 'Custom Std' lane contains Alcohol Dehydrogenase (* , 150 kDa) and Bovine Serum Albumin (** , 66 kDa). (E-F) SDS PAGE of crosslinked full-length SARS-CoV-2 N protein in 20 mM NaPi pH 7.4, 1.25mM DSS as a function of protein concentration in the presence of 200mM NaCl (E) and in the absence of added salt (F).

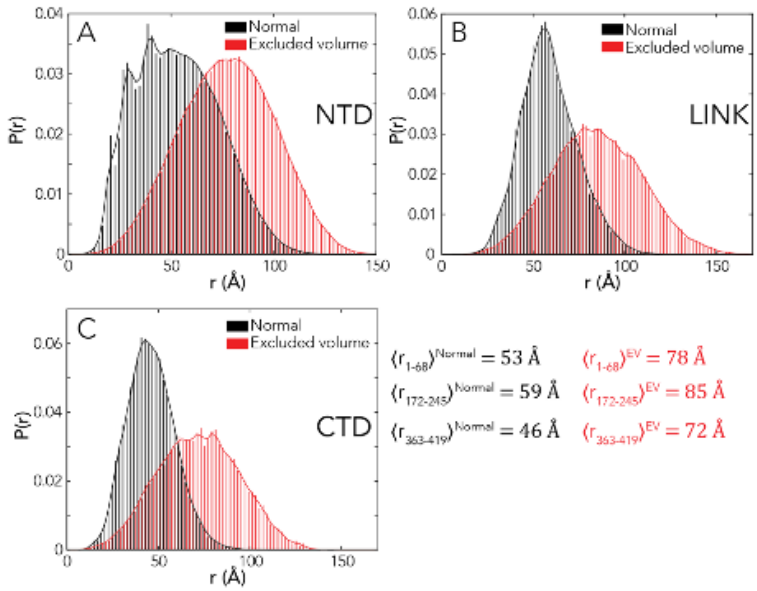


Figure D.14: Distributions of inter-residue distance from ABSINTH simulations (black) vs. excluded volume simulations (red). Comparison of simulations with the full ABSINTH Hamiltonian ('normal', black) against simulations performed in the excluded volume (EV, red) limit for **A.** NTD, **B.** LINK and **C.** CTD. In all cases the EV simulations report substantially larger average distances than the ABSINTH simulations, as expected given the absence of any attractive intramolecular interactions. The distances reported from the EV simulations are also slightly more expanded than under fully denatured conditions, consistent with systems studied previously (see [496, 514].

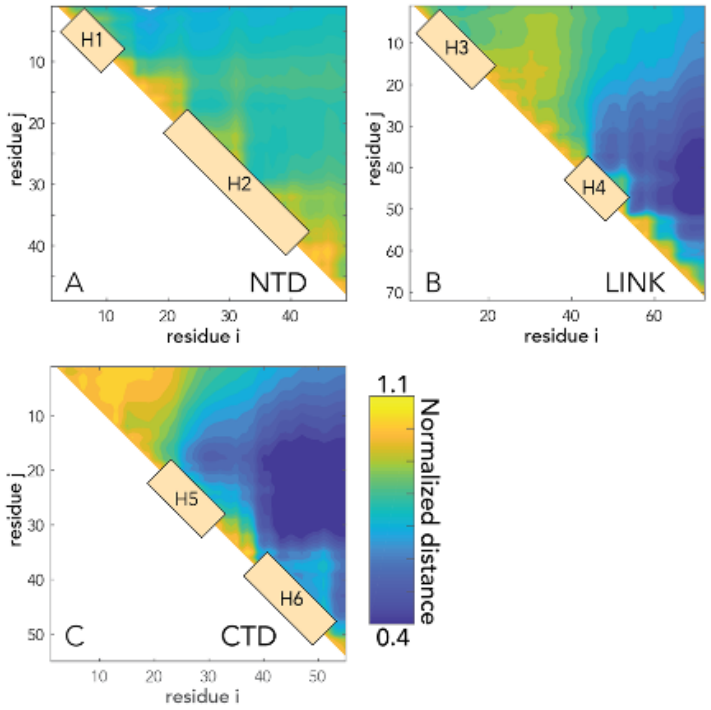


Figure D.15: Scaling maps for IDR-only simulations. Scaling maps report on the normalized distance between pairs of residues, where normalization is done by the distance expected if the IDRs behaved as self-avoiding chains in the excluded-volume limit. Scaling maps for IDR-only simulations of the **A.** NTD, **B.** LINK and **C.** CTD. For each sequence, transient helices are annotated on the scaling maps. Note that in the LINK we observe interaction between the C-terminal region of the LINK and H4, while H3 does not interact with any parts of the sequence. Similarly, in CTD we see extensive intramolecular interactions between H5 and H6.

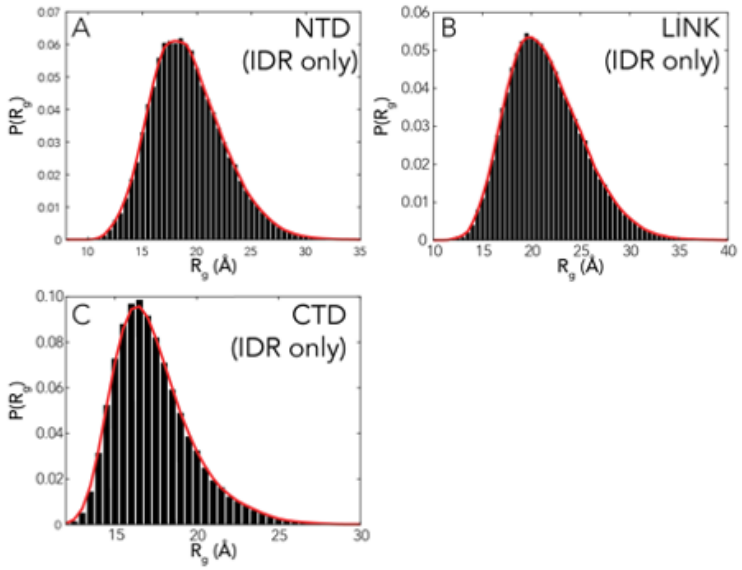


Figure D.16: Distributions for the radius of gyration (R_g) of for IDR-only simulations. R_g distributions for **A.** NTD, **B.** LINK and **C.** CTD. Average R_g for each IDR in isolation is 19.1 Å (NTD), 21.4 Å (LINK), and 17.1 Å (CTD).

Appendix E

Supplementary Material on "Antagonism between substitutions in β-lactamase explains a path not taken in the evolution of bacterial drug resistance"

The work in this appendix is published in: Brown, C.A., Hu, L., Sun, Z., Patel, M.P., Singh, S., Porter, J.R., Sankaran, B., Venkataram Prasad, B.V.V., Bowman, G.R., Palzkill, T., Antagonism between substitutions in β-lactamase explains a path not taken in the evolution of bacterial drug resistance, J.Biol., Chem., 2020, doi:10.1074/jbc.RA119.012489 [437]

E.1 Supplementary Data

Table E.1: Table S1. X-ray crystallography data collection and refinement statistics for CTX-M-14 mutant enzymes. *Values in parentheses represent the highest-resolution bin.

	P167S/D240G	E166A/D240G	E166A/P167S/ D240G	E166A/D240G-CTX	E166A/P167S/ D240G-CTX-1	E166A/P167S/ D240G-CTX-2
PDB ID	6V5E	6V6P	6V6G	6V7T	6V83	6V8V
Data collection						
Space group	P 41 21 2	P 41 21 2	P 32 2 1	P 21	P 41 21 2	P 32 2 1
a, b, c ()	42.2, 42,2, 261.6	41.9, 41.9, 259.2	41.4, 41.4, 231.1	45.1, 107.3, 47.8	42.4, 42.4, 262.7	41.3, 41.3, 231.6
α, β, γ (°)	90, 90, 90	90, 90, 90	90, 90, 120	90, 99.9, 90	90, 90, 90	90, 90, 120
Resolution Range (Å)	41.70 - 2.30	39.89 - 1.55	35.88 - 1.50	35.48 - 1.34	41.82 - 1.80	32.44 - 1.80
	(2.38 - 2.30)	(1.61 - 1.55)	(1.56 - 1.50)	(1.39 - 1.34)	(1.84 - 1.80)	(1.87 - 1.80)
R-merge (%)	8.4 (12.1)	9.7 (18.7)	5.6 (46.5)	4.7 (11.8)	9.4 (60.6)	8.7 (11.8)
I/sigma	17.4 (9.6)	16.6 (10.3)	30.8 (5.1)	17.1 (8.5)	31.3 (5.4)	11.7 (4.8)
Multiplicity	7.0 (6.4)	11.8 (14.0)	8.4 (10.6)	3.7 (3.7)	16.7 (19.9)	4.8 (2.0)
Completeness (%)	88.3	85.5	98.8	99.4	100	97.9
Wilson B-factor ($Å^2$)	29.4	10.3	12.1	9.1	18.2	14.1
Refinement						
Molecules per asymmetric unit	1	1	1	2	1	1
No. of unique reflections	10104 (880)	30051 (2829)	37869 (3494)	99523 (9924)	23526 (2299)	21924 (1871)
R-work/R-free (%)	18.5 / 25.1	16.8 / 19.4	18.3 / 21.7	14.3 / 16.1	17.1 / 20.8	14.9 / 18.9
No. of protein residues	263	263	260	526	261	260
Ramachandran						
Favored (%)	98.5	98.1	97.7	98.1	98.5	98.1
Outliers (%)	0	0	0.39	0.38	0	0.39
Average B-factor ($Å^2$)	31	14.2	21.2	14.1	23.3	16.6
Protein	30.8	12.4	19.6	11.7	21.1	14.4
Ligand	-	23.4	37.3	18.6	62.2	28.7
Solvent	34.7	29.1	32.1	26.7	36.3	30.2
RMS deviations						
Bond length ($Å$)	0.003	0.011	0.007	0.006	0.006	0.011
Bond angles (°)	0.685	1.16	0.92	0.93	0.96	1.204

E.2 Supplementary Figures

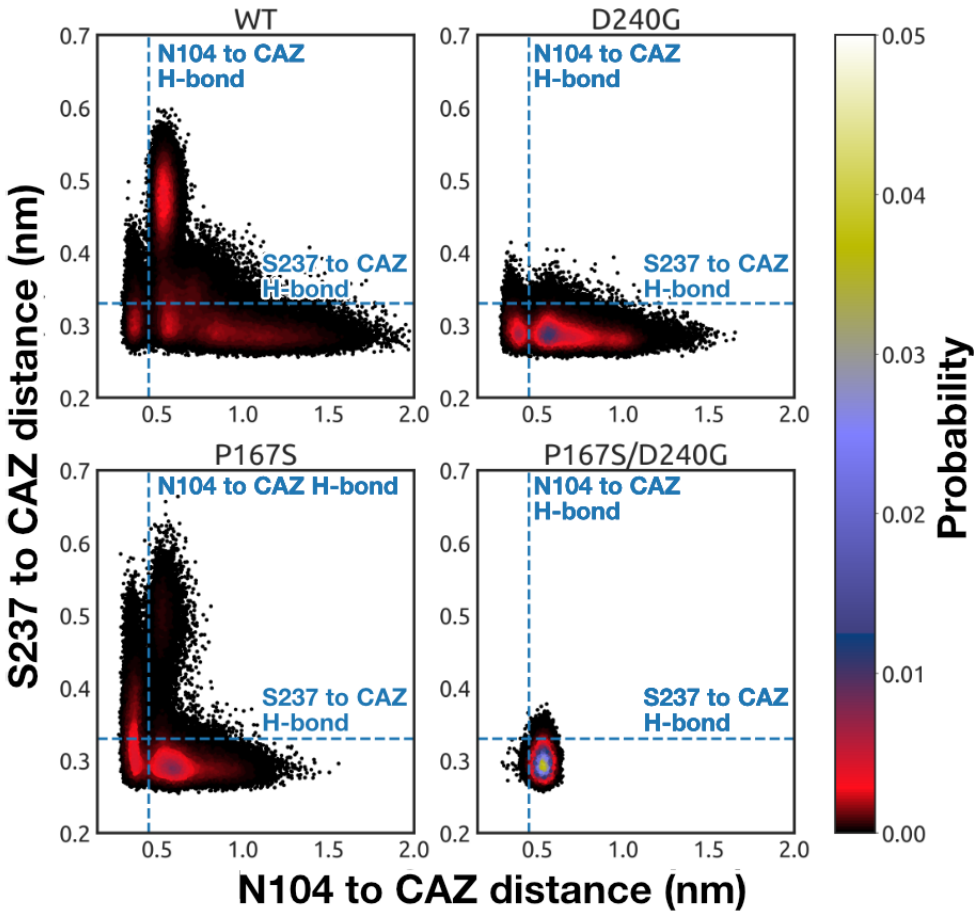


Figure E.1: The β3 loop and Asn104 contact CAZ in the single mutants. Joint distributions of two hydrogen-bonding distances that capture the contacts between CTX-M and ceftazidime (CAZ) in the acyl-enzyme complex: i) Asn104 to the imino group of ceftazidime and ii) the backbone nitrogen of S237 on the β3 loop to the β-Lactam carbonyl oxygen of ceftazidime. Distance cutoffs for hydrogen-bonding interactions are marked (dashed lines) to indicate whether or not an interaction occurs. Distributions are shown for wild type (top left), D240G (top right), P167S (bottom left), and P167S/D240G (bottom right). Each point represents a snapshot from the molecular dynamics simulations colored according to its probability based on a 2D histogram.

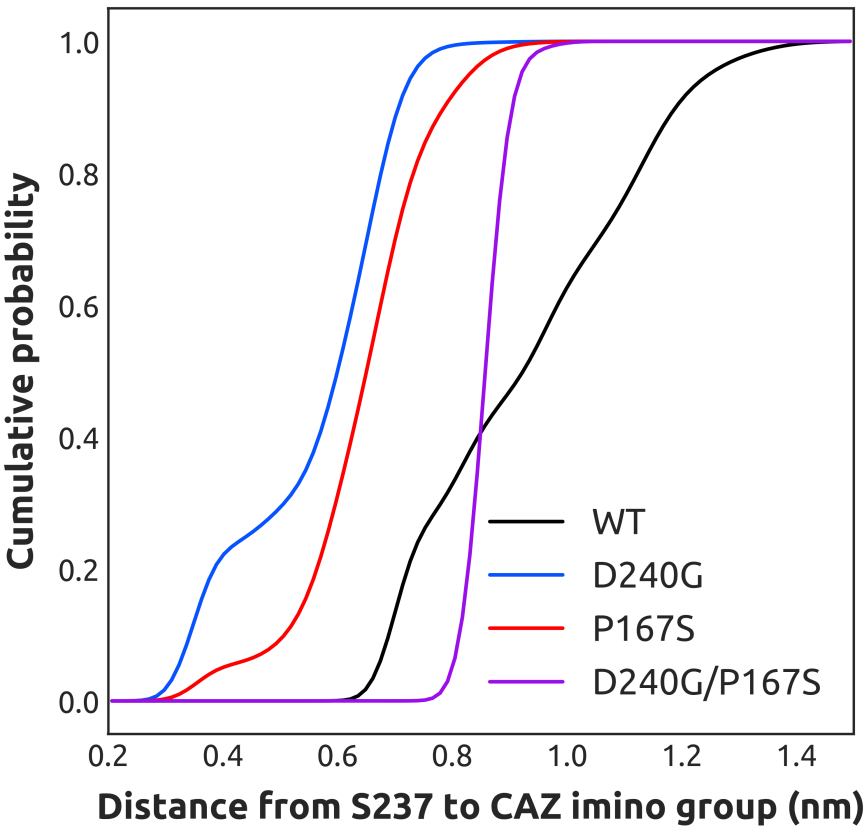


Figure E.2: Ser237 makes contacts with the imino group of ceftazidime. Cumulative distance distribution of the sidechain oxygen of Ser237 to the carboxylate of the imino group of ceftazdime in the acyl-enzyme complex. Distributions are shown for wild type (black), D240G (blue), P167S (red), and P167S/D240G (purple).

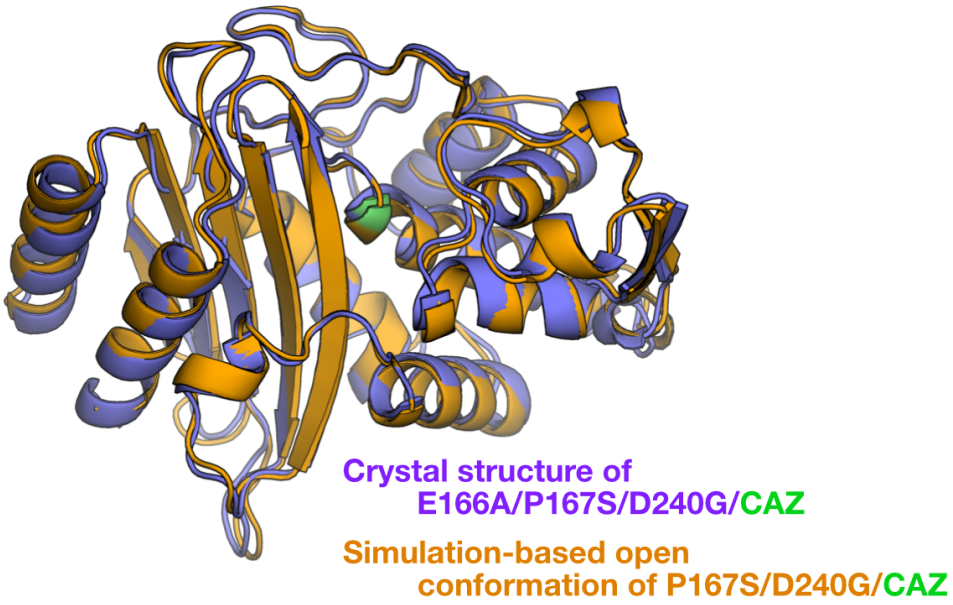


Figure E.3: MD simulations of the closed conformation of P167S/D240G capture an open conformation of the Ω-loop. Overlay of the crystal structure of E166A/P167S/D240G/CAZ (purple, Ser70 colored in green) with a representative conformation from simulation of the open conformation of the Ω-loop (orange, Ser70 colored in green) sampled from simulations of the P167S/D240G variant starting from the closed conformation. In both constructs the catalytic serine that forms the acyl-enzyme complex with the serine that binds ceftazidime (labelled CAZ) is colored green. The ceftazidime molecule is not shown for clarity.

Bibliography

[1] Andrew Waterhouse, Martino Bertoni, Stefan Bienert, Gabriel Studer, Gerardo Tauriello, Rafal Gumienny, Florian T Heer, Tjaart A P de Beer, Christine Rempfer, Lorenza Bordoli, Rosalba Lepore, and Torsten Schwede. SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Research*, 46(W1):W296–W303, July 2018.

[2] Jasmine Cubuk, Jhullian J. Alston, J. Jeremías Incicco, Sukrit Singh, Melissa D. Stuchell-Brereton, Michael D. Ward, Maxwell I. Zimmerman, Neha Vithani, Daniel Griffith, Jason A. Wagoner, Gregory R. Bowman, Kathleen B. Hall, Andrea Soranno, and Alex S. Holehouse. The SARS-CoV-2 nucleocapsid protein is dynamic, disordered, and phase separates with RNA. preprint, Biophysics, June 2020.

[3] Zhaoming Su, Chao Wu, Liuqing Shi, Priya Luthra, Grigore D. Pintilie, Britney Johnson, Justin R. Porter, Peng Ge, Muyuan Chen, Gai Liu, Thomas E. Frederick, Jennifer M. Binning, Gregory R. Bowman, Z. Hong Zhou, Christopher F. Basler, Michael L. Gross, Daisy W. Leung, Wah Chiu, and Gaya K. Amarasinghe. Electron cryo-microscopy structure of ebola virus nucleoprotein reveals a mechanism for nucleocapsid-like assembly. *Cell*, 172(5):966–978.e12, February 2018.

[4] Bryan L Roth, John J Irwin, and Brian K Shoichet. Discovery of new GPCR ligands to illuminate new biology. *Nature chemical biology*, 13(11):1143–1151, November 2017.

[5] Brian D. Weitzner, Yakov Kipnis, A. Gerard Daniel, Donald Hilvert, and David Baker. A computational method for design of connected catalytic networks in proteins. *Protein Science*, 28(12):2036–2041, December 2019.

[6] J. C. Kendrew, G. Bodo, H. M. Dintzis, R. G. Parrish, H. Wyckoff, and D. C. Phillips. A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature*, 181(4610):662–666, March 1958.

[7] Sofia Khan and Mauno Vihinen. Performance of protein stability predictors. *Human Mutation*, 31(6):675–684, March 2010.

[8] Jian Yin, Niel M. Henriksen, David R. Slochower, Michael R. Shirts, Michael W. Chiu, David L. Mobley, and Michael K. Gilson. Overview of the SAMPL5 host–guest challenge: Are we doing better? *Journal of Computer-Aided Molecular Design*, 31(1):1–19, January 2017.

[9] Kathryn M Hart, Chris M W Ho, Supratik Dutta, Michael L Gross, and Gregory R Bowman. Modelling proteins' hidden conformations to predict antibiotic resistance. *Nature communications*, 7:12965, October 2016.

[10] Justin R Porter, Artur Meller, Maxwell I Zimmerman, Michael J Greenberg, and Gregory R Bowman. Conformational distributions of isolated myosin motor domains encode their mechanochemical properties. *eLife*, 9:e55132, May 2020.

[11] C Levinthal. How to Fold Graciously. *Topics in mossbauer spectroscopy*, 20(1):25–44, October 1969.

[12] Robert L. Baldwin. Early days of protein hydrogen exchange: 1954-1972. *Proteins: Structure, Function, and Bioinformatics*, 79(7):2021–2026, July 2011.

[13] Katherine A Henzler-Wildman, Ming Lei, Vu Thai, S Jordan Kerns, Martin Karplus, and Dorothee Kern. A hierarchy of timescales in protein dynamics is linked to enzyme catalysis. *Nature*, 450(7171):913–916, December 2007.

[14] David D Boehr, Dan McElheny, H Jane Dyson, and P E Wright. The dynamic energy landscape of dihydrofolate reductase catalysis. *Science*, 313(5793):1638–1642, September 2006.

[15] David D Boehr, Ruth Nussinov, and P E Wright. The role of dynamic conformational ensembles in biomolecular recognition. *Nature chemical biology*, 5(11):789–796, November 2009.

[16] Adelajda Zorba, Vanessa Buosi, Steffen Kutter, Nadja Kern, Francesco Pontiggia, Young-Jin Cho, and Dorothee Kern. Molecular mechanism of Aurora A kinase autophosphorylation and its allosteric activation by TPX2. *eLife*, 3:e02667, May 2014.

[17] S Jordan Kerns, Roman V Agafonov, Young-Jin Cho, Francesco Pontiggia, Renee Otten, Dimitar V Pachov, Steffen Kutter, Lien A Phung, Padraig N Murphy, Vu Thai, Tom Alber, Michael F Hagan, and Dorothee Kern. The energy landscape of adenylate kinase during catalysis. *Nature Structural & Molecular Biology*, 22(2):124–131, February 2015.

[18] Ned Van Eps, Lori L Anderson, Oleg G Kisselev, Thomas J Baranski, Wayne L Hubbell, and Garland R Marshall. Electron paramagnetic resonance studies of functionally active, nitroxide spin-labeled peptide analogues of the C-terminus of a G-protein alpha subunit. *Biochemistry*, 49(32):6877–6886, August 2010.

[19] A Joshua Wand. The dark energy of proteins comes to light: conformational entropy and its role in protein function revealed by NMR relaxation. 23(1):75–81, February 2013.

[20] Antoine Koehl, Hongli Hu, Shoji Maeda, Yan Zhang, Qianhui Qu, Joseph M Paggi, Naomi R Latorraca, Daniel Hilger, Roger Dawson, Hugues Matile, Gebhard F X Schertler, Sébastien Granier, William I Weis, Ron O Dror, Aashish Manglik, Georgios Skiniotis, and Brian K Kobilka. Structure of the μ-opioid receptor–G i protein complex. *Nature*, 383:1, June 2018.

[21] Christopher J Draper-Joyce, Maryam Khoshouei, David M Thal, Yi-Lynn Liang, Anh T N Nguyen, Sebastian G B Furness, Hariprasad Venugopal, Jo-Anne Baltos, Jürgen M Plitzko, Radostin Danev, Wolfgang Baumeister, Lauren T May, Denise Wootten, Patrick M Sexton, Alisa Glukhova, and Arthur Christopoulos. Structure of the adenosine-bound human adenosine A1 receptor-Gi complex. *Nature*, 63:1, June 2018.

[22] Javier García-Nafría, Rony Nehmé, Patricia C Edwards, and Christopher G Tate. Cryo-EM structure of the serotonin 5-HT1B receptor coupled to heterotrimeric Go. *Nature*, 7:118, June 2018.

[23] Yanyong Kang, Oleg Kuybeda, Parker W de Waal, Somnath Mukherjee, Ned Van Eps, Przemyslaw Dutka, X Edward Zhou, Alberto Bartesaghi, Satchal Erramilli, Takefumi Morizumi, Xin Gu, Yanting Yin, Ping Liu, Yi Jiang, Xing Meng, Gongpu Zhao, Karsten Melcher, Oliver P Ernst, Anthony A Kossiakoff, Sriram Subramaniam, and H Eric Xu. Cryo-EM structure of human rhodopsin bound to an inhibitory G protein. *Nature*, 63(Suppl. 1):1256, June 2018.

[24] Ron O Dror, Robert M Dirks, J P Grossman, Huafeng Xu, and David E Shaw. Biomolecular Simulation: A Computational Microscope for Molecular Biology. *dx.doi.org*, 41(1):429–452, May 2012.

[25] Eric H. Lee, Jen Hsin, Marcos Sotomayor, Gemma Comellas, and Klaus Schulten. Discovery through the computational microscope. *Structure*, 17(10):1295–1306, October 2009.

[26] Pedro E. M. Lopes, Olgun Guvench, and Alexander D. MacKerell. Current status of protein force fields for molecular dynamics simulations. In Andreas Kukol, editor, *Molecular Modeling of Proteins*, volume 1215, pages 47–71. Springer New York, New York, NY, 2015.

[27] Weinan E and Eric Vanden-Eijnden. Transition-Path Theory and Path-Finding Algorithms for the Study of Rare Events. *dx.doi.org*, 61(1):391–420, March 2010.

[28] Donald Hamelberg, John Mongan, and J. Andrew McCammon. Accelerated molecular dynamics: A promising and efficient simulation method for biomolecules. *The Journal of Chemical Physics*, 120(24):11919–11929, June 2004.

[29] Dietmar Paschek, Hugh Nymeyer, and Angel E. García. Replica exchange simulation of reversible folding/unfolding of the Trp-cage miniprotein in explicit solvent: On the structure and possible role of internal water. *Journal of Structural Biology*, 157(3):524–533, March 2007.

[30] Stefano Piana and Alessandro Laio. A bias-exchange approach to protein folding. *The Journal of Physical Chemistry B*, 111(17):4553–4559, May 2007.

[31] John E. Stone, James C. Phillips, Peter L. Freddolino, David J. Hardy, Leonardo G. Trabuco, and Klaus Schulten. Accelerating molecular modeling applications with graphics processors. *Journal of Computational Chemistry*, 28(16):2618–2640, December 2007.

[32] Mark S. Friedrichs, Peter Eastman, Vishal Vaidyanathan, Mike Houston, Scott Legrand, Adam L. Beberg, Daniel L. Ensign, Christopher M. Bruns, and Vijay S. Pande. Accelerating molecular dynamic simulation on graphics processing units. *Journal of Computational Chemistry*, 30(6):864–872, April 2009.

[33] Peter Eastman and Vijay Pande. Openmm: a hardware-independent framework for molecular simulations. *Computing in Science & Engineering*, 12(4):34–39, July 2010.

[34] Peter Eastman, Jason Swails, John D. Chodera, Robert T. McGibbon, Yutong Zhao, Kyle A. Beauchamp, Lee-Ping Wang, Andrew C. Simmonett, Matthew P. Harrigan, Chaya D. Stern, Rafal P. Wiewiora, Bernard R. Brooks, and Vijay S. Pande. OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLOS Computational Biology*, 13(7):e1005659, July 2017.

[35] Maxwell I. Zimmerman and Gregory R. Bowman. Fast conformational searches by balancing exploration/exploitation trade-offs. *Journal of Chemical Theory and Computation*, 11(12):5747–5757, December 2015.

[36] David E. Shaw, Martin M. Deneroff, Ron O. Dror, Jeffrey S. Kuskin, Richard H. Larson, John K. Salmon, Cliff Young, Brannon Batson, Kevin J. Bowers, Jack C. Chao, Michael P. Eastwood, Joseph Gagliardo, J. P. Grossman, C. Richard Ho, Douglas J. Ierardi, István Kolossváry, John L. Klepeis, Timothy Layman, Christine McLeavey, Mark A. Moraes, Rolf Mueller, Edward C. Priest, Yibing Shan, Jochen Spengler, Michael Theobald, Brian Towles, and Stanley C. Wang. Anton, a special-purpose machine for molecular dynamics simulation. *Communications of the ACM*, 51(7):91–97, July 2008.

[37] M Shirts and V S Pande. COMPUTING: Screen Savers of the World Unite! *Science*, 290(5498):1903–1904, December 2000.

[38] Brooke E Husic and Vijay S Pande. Markov State Models: From an Art to a Science. *Journal of the American Chemical Society*, page jacs.7b12191, January 2018.

[39] J. R. Porter, M. I. Zimmerman, and G. R. Bowman. Enspara: Modeling molecular ensembles with scalable data structures and parallel computing. *The Journal of Chemical Physics*, 150(4):044108, 2019.

[40] John D Chodera and Frank Noé. Markov state models of biomolecular conformational dynamics. *Current opinion in structural biology*, 25:135–144, April 2014.

[41] Catherine R Knoverek, Gaya K Amarasinghe, and Gregory R Bowman. Advanced Methods for Accessing Protein Shape-Shifting Present New Therapeutic Opportunities. *Trends in biochemical sciences*, December 2018.

[42] Maxwell I Zimmerman, Justin R Porter, Xianqiang Sun, Roseane R Silva, and Gregory R Bowman. Choice of Adaptive Sampling Strategy Impacts State Discovery, Transition Probabilities, and the Apparent Mechanism of Conformational Changes. *Journal of Chemical Theory and Computation*, q-bio.BM:acs.jctc.8b00500, October 2018.

[43] L Molgedey and HG Schuster. Separation of a mixture of independent signals using time delayed correlations. *Physical review letters*, 72(23):3634–3637, June 1994.

[44] Mohammad M Sultan and Vijay S Pande. tICA-Metadynamics: Accelerating Metadynamics by Using Kinetically Selected Collective Variables. *Journal of Chemical Theory and Computation*, 13(6):2440–2447, June 2017.

[45] Guillermo Pérez-Hernández, Fabian Paul, Toni Giorgino, Gianni De Fabritiis, and Frank Noé. Identification of slow molecular order parameters for Markov model construction. *arXiv.org*, (1):015102, February 2013.

[46] Vincent A Voelz, Gregory R Bowman, Kyle Beauchamp, and Vijay S Pande. Molecular simulation of ab initio protein folding for a millisecond folder NTL9(1-39). *Journal of the American Chemical Society*, 132(5):1526–1528, February 2010.

[47] Gregory R Bowman and Phillip L Geissler. Equilibrium fluctuations of a single folded protein reveal a multitude of potential cryptic allosteric sites. *Proceedings of the National Academy of Sciences*, 109(29):11681–11686, July 2012.

[48] Xianqiang Sun, Sukrit Singh, Kendall Blumer, and Gregory R Bowman. Simulation of spontaneous G protein activation reveals a new intermediate driving GDP unbinding. *eLife*, 7, October 2018.

[49] Justin R Porter, Katelyn E Moeder, Carrie A Sibbald, Maxwell I Zimmerman, Kathryn M Hart, Michael J Greenberg, and Gregory R Bowman. Cooperative Changes in Solvent Exposure Identify Cryptic Pockets, Switches, and Allosteric Coupling. *Biophysical Journal*, 116(5):818–830, March 2019.

[50] Teofilo F. Gonzalez. Clustering to minimize the maximum intercluster distance. *Theoretical Computer Science*, 38:293–306, 1985.

[51] Hae-Sang Park and Chi-Hyuck Jun. A simple and fast algorithm for K-medoids clustering. *Expert Systems with Applications*, 36(2):3336–3341, March 2009.

[52] Jan-Hendrik Prinz, John D. Chodera, Vijay S. Pande, William C. Swope, Jeremy C. Smith, and Frank Noé. Optimal use of data in parallel tempering simulations for the

construction of discrete-state Markov models of biomolecular dynamics. *The Journal of Chemical Physics*, 134(24):244108, June 2011.

[53] Jan-Hendrik Prinz, Hao Wu, Marco Sarich, Bettina Keller, Martin Senne, Martin Held, John D. Chodera, Christof Schütte, and Frank Noé. Markov models of molecular kinetics: Generation and validation. *The Journal of Chemical Physics*, 134(17):174105, May 2011.

[54] Matthew A. Cruz, Thomas E. Frederick, Sukrit Singh, Neha Vithani, Maxwell I. Zimmerman, Justin R. Porter, Katelyn E. Moeder, Gaya K. Amarasinghe, and Gregory R. Bowman. Discovery of a cryptic allosteric site in ebola's 'undruggable' vp35 protein using simulations and experiments. *bioRxiv*, 2020.

[55] Gregory R Bowman, Vincent A Voelz, and Vijay S Pande. Taming the complexity of protein folding. 21(1):4–11, February 2011.

[56] Gregory R Bowman. Accurately modeling nanosecond protein dynamics requires at least microseconds of simulation. *Journal of computational chemistry*, pages n/a–n/a, June 2015.

[57] Gregory R Bowman and Phillip L Geissler. Extensive conformational heterogeneity within protein cores. *The Journal of Physical Chemistry B*, 118(24):6417–6423, 2014.

[58] Yunhui Ge, Elias Borne, Shannon Stewart, Michael R. Hansen, Emilia C. Arturo, Eileen K. Jaffe, and Vincent A. Voelz. Simulations of the regulatory ACT domain of human phenylalanine hydroxylase (Pah) unveil its mechanism of phenylalanine binding. *Journal of Biological Chemistry*, 293(51):19532–19543, December 2018.

[59] Shi Chen, Rafal P Wiewiora, Fanwang Meng, Nicolas Babault, Anqi Ma, Wenyu Yu, Kun Qian, Hao Hu, Hua Zou, Junyi Wang, Shijie Fan, Gil Blum, Fabio Pittella-Silva, Kyle A Beauchamp, Wolfram Tempel, Hualiang Jiang, Kaixian Chen, Robert J Skene, Yujun George Zheng, Peter J Brown, Jian Jin, Cheng Luo, John D Chodera, and Minkui

Luo. The dynamic conformational landscape of the protein methyltransferase SETD8. *eLife*, 8:e45403, May 2019.

[60] Roberta Pascolutti, Xianqiang Sun, Joseph Kao, Roy L. Maute, Aaron M. Ring, Gregory R. Bowman, and Andrew C. Kruse. Structure and dynamics of pd-l1 and an ultra-high-affinity pd-1 receptor mutant. *Structure*, 24(10):1719–1728, October 2016.

[61] V. J. Hilser. An ensemble view of allostery. *Science*, 327(5966):653–654, February 2010.

[62] M F Perutz. Stereochemistry of cooperative effects in haemoglobin. *Nature*, 228(5273):726–739, November 1970.

[63] Shiou-Ru Tzeng and Charalampos G Kalodimos. Dynamic activation of an allosteric regulatory protein. *Nature*, 462(7271):368–372, November 2009.

[64] Shiou-Ru Tzeng and Charalampos G Kalodimos. Protein dynamics and allostery: an NMR view. 21(1):62–67, February 2011.

[65] William I Weis and Brian K Kobilka. The Molecular Basis of G Protein–Coupled Receptor Activation. *Annual review of biochemistry*, 87(1):897–919, June 2018.

[66] Daniel Wacker, Raymond C. Stevens, and Bryan L. Roth. How ligands illuminate gpcr molecular pharmacology. *Cell*, 170(3):414–427, July 2017.

[67] Kadla R Rosholm, Natascha Leijnse, Anna Mantsiou, Vadym Tkach, Søren L Pedersen, Volker F Wirth, Lene B Oddershede, Knud J Jensen, Karen L Martinez, Nikos S Hatzakis, Poul Martin Bendix, Andrew Callan-Jones, and Dimitrios Stamou. Membrane curvature regulates ligand-specific membrane sorting of GPCRs in living cells. *Nature Chemical Biology*, 13(7):724–729, July 2017.

[68] Søren G F Rasmussen, Brian T DeVree, Yaozhong Zou, Andrew C Kruse, Ka Young Chung, Tong Sun Kobilka, Foon Sun Thian, Pil Seok Chae, Els Pardon, Diane Calinski,

Jesper M Mathiesen, Syed T A Shah, Joseph A Lyons, Martin Caffrey, Samuel H Gellman, Jan Steyaert, Georgios Skiniotis, William I Weis, Roger K Sunahara, and Brian K Kobilka. Crystal structure of the β2 adrenergic receptor-Gs protein complex. *Nature*, 477(7366):549–555, July 2011.

[69] K Gunasekaran, Buyong Ma, and Ruth Nussinov. Is allostery an intrinsic property of all dynamic proteins? *Proteins*, 57(3):433–443, November 2004.

[70] Philip A. Romero and Frances H. Arnold. Exploring protein fitness landscapes by directed evolution. *Nature Reviews Molecular Cell Biology*, 10(12):866–876, December 2009.

[71] Merijn L M Salverda, J Arjan G M De Visser, and Miriam Barlow. Natural evolution of TEM-1 β-lactamase: experimental reconstruction and clinical relevance. *FEMS microbiology reviews*, 34(6):1015–1036, November 2010.

[72] Michelle R. Arkin and James A. Wells. Small-molecule inhibitors of protein–protein interactions: progressing towards the dream. *Nature Reviews Drug Discovery*, 3(4):301–317, April 2004.

[73] Anthony Ivetac and J. Andrew McCammon. Mapping the druggable allosteric space of g-protein coupled receptors: a fragment-based molecular dynamics approach: computational mapping of novel druggable sites on gpcrs. *Chemical Biology & Drug Design*, pages no–no, July 2010.

[74] Jeanne A Hardy and James A Wells. Searching for new allosteric sites in enzymes. *Current Opinion in Structural Biology*, 14(6):706–715, December 2004.

[75] James R Horn and Brian K Shoichet. Allosteric Inhibition Through Core Disruption. *Journal of Molecular Biology*, 336(5):1283–1291, March 2004.

[76] M. R. Arkin, M. Randal, W. L. DeLano, J. Hyde, T. N. Luong, J. D. Oslob, D. R. Raphael, L. Taylor, J. Wang, R. S. McDowell, J. A. Wells, and A. C. Braisted. Binding

of small molecules to an adaptive protein-protein interface. *Proceedings of the National Academy of Sciences*, 100(4):1603–1608, February 2003.

[77] Jonathan M. Ostrem, Ulf Peters, Martin L. Sos, James A. Wells, and Kevan M. Shokat. K-Ras(G12c) inhibitors allosterically control GTP affinity and effector interactions. *Nature*, 503(7477):548–551, November 2013.

[78] Sandor Vajda, Dmitri Beglov, Amanda E Wakefield, Megan Egbert, and Adrian Whitty. Cryptic binding sites on proteins: definition, detection, and druggability. *Current opinion in chemical biology*, 44:1–8, May 2018.

[79] Julie R. Schames, Richard H. Henchman, Jay S. Siegel, Christoph A. Sotriffer, Haihong Ni, and J. Andrew McCammon. Discovery of a novel binding trench in hiv integrase. *Journal of Medicinal Chemistry*, 47(8):1879–1881, April 2004.

[80] Patrick A. Frantom, Hui-Min Zhang, Mark R. Emmett, Alan G. Marshall, and John S. Blanchard. Mapping of the allosteric network in the regulation of -isopropylmalate synthase from *mycobacterium tuberculosis* by the feedback inhibitor l -leucine: solution-phase h/d exchange monitored by ft-icr mass spectrometry. *Biochemistry*, 48(31):7457–7464, August 2009.

[81] Gregory Manley and J. Patrick Loria. NMR insights into protein allostery. *Archives of Biochemistry and Biophysics*, 519(2):223–231, March 2012.

[82] S W Lockless and R Ranganathan. Evolutionarily conserved pathways of energetic connectivity in protein families. *Science*, 286(5438):295–299, October 1999.

[83] Victoria A Feher, Jacob D Durrant, Adam T Van Wart, and Rommie E Amaro. Computational approaches to mapping allosteric pathways. 25:98–103, April 2014.

[84] Joe G Greener and Michael Je Sternberg. Structure-based prediction of protein allostery. 50:1–8, October 2017.

[85] Toshiko Ichiye and Martin Karplus. Collective motions in proteins: A covariance analysis of atomic fluctuations in molecular dynamics and normal mode simulations. *Proteins*, 11(3):205–217, November 1991.

[86] Christopher L McClendon, Gregory Friedland, David L Mobley, Homeira Amirkhani, and Matthew P Jacobson. Quantifying Correlations Between Allosteric Sites in Thermodynamic Ensembles. *Journal of Chemical Theory and Computation*, 5(9):2486–2502, September 2009.

[87] A Cooper and D T Dryden. Allostery without conformational change. A plausible model. *European biophysics journal : EBJ*, 11(2):103–109, 1984.

[88] Nataliya Popovych, Shangjin Sun, Richard H Ebright, and Charalampos G Kalodimos. Dynamically driven protein allostery. *Nature Structural & Molecular Biology*, 13(9):831–838, September 2006.

[89] Milo M Lin. Timing Correlations in Proteins Predict Functional Modules and Dynamic Allostery. *Journal of the American Chemical Society*, page jacs.5b08814, April 2016.

[90] Sukrit Singh and Gregory R Bowman. Quantifying allosteric communication via both concerted structural changes and conformational disorder with CARDS. *Journal of Chemical Theory and Computation*, page acs.jctc.6b01181, March 2017.

[91] Daniel M Rosenbaum, Søren G F Rasmussen, and Brian K Kobilka. The structure and function of G-protein-coupled receptors. *Nature*, 459(7245):356–363, May 2009.

[92] Brian K Shoichet and Brian K Kobilka. Structure-based drug screening for G-protein-coupled receptors. *Trends in pharmacological sciences*, 33(5):268–272, May 2012.

[93] Jingjing Guo and Huan-Xiang Zhou. Protein Allostery and Conformational Dynamics. *Chemical Reviews*, 116(11):6503–6515, February 2016.

[94] D E Koshland. Enzyme flexibility and enzyme action. *Journal of Cellular Physiology*, 54(S1):245–258, December 1959.

[95] D E Koshland Jr, G Nemethy, and D Filmer. Comparison of Experimental Binding Data and Theoretical Models in Proteins Containing Subunits *. *Biochemistry*, 5(1):365–385, January 1966.

[96] Jacque Monod, Jeffries Wyman, and Jean-Pierre Changeux. On the nature of allosteric transitions: A plausible model. *Journal of Molecular Biology*, 12(1):88–118, May 1965.

[97] Jean-Pierre Changeux and Stuart Edelstein. Conformational selection or induced fit? 50 years of debate resolved. *F1000 biology reports*, 3(19):19, 2011.

[98] Peter Csermely, Robin Palotai, and Ruth Nussinov. Induced fit, conformational selection and independent dynamic segments: an extended view of binding events. *Trends in biochemical sciences*, 35(10):539–546, October 2010.

[99] Daniel-Adriano Silva, Gregory R Bowman, Alejandro Sosa-Peinado, and Xuhui Huang. A role for both conformational selection and induced fit in ligand binding by the LAO protein. *PLoS Computational Biology*, 7(5):e1002054, May 2011.

[100] Gordon G Hammes, Yu-Chu Chang, and Terrence G Oas. Conformational selection or induced fit: a flux description of reaction mechanism. *Proceedings of the National Academy of Sciences of the United States of America*, 106(33):13737–13741, August 2009.

[101] Jennifer L Radkiewicz and and III Charles L Brooks. Protein Dynamics in Enzymatic Catalysis: Exploration of Dihydrofolate Reductase. *Journal of the American Chemical Society*, 122(2):225–231, December 1999.

[102] Pratul K Agarwal, Salomon R Billeter, , and Sharon Hammes-Schiffer. *Nuclear Quantum Effects and Enzyme Dynamics in Dihydrofolate Reductase Catalysis*, volume 106. American Chemical Society, February 2002.

[103] Pratul K Agarwal, Salomon R Billeter, P T Ravi Rajagopalan, Stephen J Benkovic, and Sharon Hammes-Schiffer. Network of coupled promoting motions in enzyme catalysis. *Proceedings of the National Academy of Sciences*, 99(5):2794–2799, March 2002.

[104] Thomas H Rod, Jennifer L Radkiewicz, and Charles L Brooks III. Correlated motion and the effect of distal mutations in dihydrofolate reductase. *Proceedings of the National Academy of Sciences*, 100(12):6980–6985, June 2003.

[105] Oliver F Lange and Helmut Grubmüller. Generalized correlation for biomolecular dynamics. *Proteins*, 62(4):1053–1061, March 2006.

[106] Matthew J Whitley and Andrew L Lee. Frameworks for understanding long-range intra-protein communication. *Current protein & peptide science*, 10(2):116–127, April 2009.

[107] Patrick Weinkam, Jaume Pons, and Andrej Sali. Structure-based model of allostery predicts coupling between distant sites. *Proceedings of the National Academy of Sciences of the United States of America*, 109(13):4875–4880, March 2012.

[108] Tom Lenaerts, Jesper Ferkinghoff-Borg, Francois Stricher, Luis Serrano, Joost W H Schymkowitz, and Frederic Rousseau. Quantifying information transfer by protein domains: analysis of the Fyn SH2 domain structure. *BMC structural biology*, 8(1):43, 2008.

[109] Kateri H Dubay, Jacques P Bothma, and Phillip L Geissler. Long-range intra-protein communication can be transmitted by correlated side-chain fluctuations alone. *PLoS Computational Biology*, 7(9):e1002168, September 2011.

[110] R Bryn Fenwick, Laura Orellana, Santi Esteban-Martín, Modesto Orozco, and Xavier Salvatella. Correlated motions are a fundamental property of β-sheets. *Nature communications*, 5:4070, June 2014.

[111] Qiang Cui and Martin Karplus. Allostery and cooperativity revisited. *Protein Science*, 17(8):1295–1307, August 2008.

[112] Ron Elber. Simulations of allosteric transitions. 21(2):167–172, April 2011.

[113] Adam T Van Wart, Jacob Durrant, Lane Votapka, and Rommie E Amaro. Weighted Implementation of Suboptimal Paths (WISP): An Optimized Algorithm and Tool for Dynamical Network Analysis. *Journal of Chemical Theory and Computation*, 10(2):511–517, February 2014.

[114] Robert D Malmstrom, Alexandr P Kornev, Susan S Taylor, and Rommie E Amaro. Allostery through the computational microscope: cAMP activation of a canonical signalling domain. *Nature communications*, 6:7588, July 2015.

[115] Jhih-Wei Chu and Gregory A Voth. Allostery of actin filaments: molecular dynamics simulations and coarse-grained analysis. *Proceedings of the National Academy of Sciences*, 102(37):13111–13116, September 2005.

[116] Krishna Pratap Ravindranathan, Emilio Gallicchio, and Ronald M Levy. Conformational equilibria and free energy profiles for the allosteric transition of the ribose-binding protein. *Journal of Molecular Biology*, 353(1):196–210, October 2005.

[117] Patrick Weinkam, Yao Chi Chen, Jaume Pons, and Andrej Sali. Impact of mutations on the allosteric conformational equilibrium. *Journal of Molecular Biology*, 425(3):647–661, February 2013.

[118] Ying Liu and Ivet Bahar. Sequence evolution correlates with structural dynamics. *Molecular biology and evolution*, 29(9):2253–2263, September 2012.

[119] Liqun Zhang, Sabine Bouguet-Bonnet, and Matthias Buck. *Combining NMR and Molecular Dynamics Studies for Insights into the Allostery of Small GTPase–Protein Interactions*, pages 235–259. Springer New York, New York, NY, 2012.

[120] Dorothee Kern and Erik RP Zuiderweg. The role of dynamics in allosteric regulation. 13(6):748–757, December 2003.

[121] Vincent J Hilser and E Brad Thompson. Intrinsic disorder as a mechanism to optimize allosteric coupling in proteins. *Proceedings of the National Academy of Sciences*, 104(20):8311–8315, May 2007.

[122] Virginia M Burger, Diego O Nolasco, and Collin M Stultz. Expanding the Range of Protein Function at the Far End of the Order-Structure Continuum. *Journal of Biological Chemistry*, 291(13):6706–6713, March 2016.

[123] Turkan Haliloglu and Ivet Bahar. Adaptability of protein structures to enable functional interactions and evolutionary implications. 35:17–23, December 2015.

[124] Katherine Henzler-Wildman and Dorothee Kern. Dynamic personalities of proteins. *Nature*, 450(7172):964–972, December 2007.

[125] Vincent J Hilser, James O Wrabl, and Hesam N Motlagh. Structural and energetic basis of allostery. *Annual review of biophysics*, 41(1):585–609, 2012.

[126] Ursula Jakob, Richard Kriwacki, and Vladimir N Uversky. Conditionally and transiently disordered proteins: awakening cryptic disorder to regulate protein function. *Chemical Reviews*, 114(13):6779–6805, July 2014.

[127] James G Harman. Allosteric regulation of the cAMP receptor protein. *Biochimica et Biophysica Acta (BBA) - Protein Structure and Molecular Enzymology*, 1547(1):1–17, May 2001.

[128] T Heyduk and J C Lee. Escherichia coli cAMP receptor protein: evidence for three protein conformational states with different promoter binding affinities. *Biochemistry*, 28(17):6914–6924, August 1989.

[129] S C Schultz, G C Shields, and T A Steitz. Crystal structure of a CAP-DNA complex: the DNA is bent by 90 degrees. *Science*, 253(5023):1001–1007, August 1991.

[130] Seung-Hyeon Seok, Hookang Im, Hyung-Sik Won, Min-Duk Seo, Yoo-Sup Lee, Hye-Jin Yoon, Min-Jeong Cha, Jin-Young Park, and Bong-Jin Lee. Structures of inactive

CRP species reveal the atomic details of the allosteric transition that discriminates cyclic nucleotide second messengers. *Acta crystallographica. Section D, Biological crystallography*, 70(Pt 6):1726–1742, June 2014.

[131] Nataliya Popovych, Shiou-Ru Tzeng, Marco Tonelli, Richard H Ebright, and Charalampos G Kalodimos. Structural basis for cAMP-mediated allosteric control of the catabolite activator protein. *Proceedings of the National Academy of Sciences of the United States of America*, 106(17):6927–6932, April 2009.

[132] Shiou-Ru Tzeng and Charalampos G Kalodimos. Protein activity regulation by conformational entropy. *Nature*, 488(7410):236–240, August 2012.

[133] V J Hilser, D Dowdy, T G Oas, and E Freire. The structural distribution of cooperative interactions in proteins: analysis of the native state ensemble. *Proceedings of the National Academy of Sciences*, 95(17):9903–9908, August 1998.

[134] V J Hilser and E Freire. Structure-based calculation of the equilibrium folding pathway of proteins. Correlation with hydrogen exchange protection factors. *Journal of Molecular Biology*, 262(5):756–772, October 1996.

[135] Enrique Marcos, Ramon Crehuet, and Ivet Bahar. Changes in dynamics upon oligomerization regulate substrate binding and allostery in amino acid kinase family members. *PLoS Computational Biology*, 7(9):e1002201, September 2011.

[136] Gürol M Süel, Steve W Lockless, Mark A Wall, and Rama Ranganathan. Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nature structural biology*, 10(1):59–69, January 2003.

[137] M D Ediger. Spatially heterogeneous dynamics in supercooled liquids. *Annual review of physical chemistry*, 51(1):99–128, 2000.

[138] Sharon C Glotzer. Spatially heterogeneous dynamics in liquids: insights from simulation. *Journal of Non-Crystalline Solids*, 274(1-3):342–355, September 2000.

[139] Ranko Richert. Heterogeneous dynamics in liquids: fluctuations in space and time. *Journal of Physics: Condensed Matter*, 14(23):R703–R738, June 2002.

[140] Mauro Merolle, Juan P Garrahan, and David Chandler. Space-time thermodynamics of the glass transition. *Proceedings of the National Academy of Sciences*, 102(31):10837–10840, August 2005.

[141] Lester O Hedges, Lutz Maibaum, David Chandler, and Juan P Garrahan. Decoupling of exchange and persistence times in atomistic models of glass formers. 127(21):211101, December 2007.

[142] Aaron S Keys, Lester O Hedges, Juan P Garrahan, Sharon C Glotzer, and David Chandler. Excitations Are Localized and Relaxation Is Hierarchical in Glass-Forming Liquids. *Physical Review X*, 1(2):021013, November 2011.

[143] William L Jorgensen, Jayaraman Chandrasekhar, Jeffry D Madura, Roger W Impey, and Michael L Klein. Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics*, 79(2):926–935, July 1983.

[144] W L DeLano. *The PyMOL Molecular Graphics System, Ver. 1.3*. Schrödinger, 2010.

[145] David Van Der Spoel, Erik Lindahl, Berk Hess, Gerrit Groenhof, Alan E Mark, and Herman J C Berendsen. GROMACS: fast, flexible, and free. *Journal of computational chemistry*, 26(16):1701–1718, December 2005.

[146] Yong Duan, Chun Wu, Shibasish Chowdhury, Mathew C Lee, Guoming Xiong, Wei Zhang, Rong Yang, Piotr Cieplak, Ray Luo, Taisung Lee, James Caldwell, Junmei Wang, and Peter Kollman. A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *Journal of computational chemistry*, 24(16):1999–2012, December 2003.

[147] Gregory R Bowman, Eric R Bolin, Kathryn M Hart, Brendan C Maguire, and Susan Marqusee. Discovery of multiple hidden allosteric sites by combining Markov state

models and experiments. *Proceedings of the National Academy of Sciences of the United States of America*, 112(9):2734–2739, March 2015.

[148] Robert T McGibbon, Kyle A Beauchamp, Matthew P Harrigan, Christoph Klein, Jason M Swails, Carlos X Hernández, Christian R Schwantes, Lee-Ping Wang, Thomas J Lane, and Vijay S Pande. MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophysical Journal*, 109(8):1528–1532, October 2015.

[149] Gerhard Hummer. From transition paths to transition states and rate coefficients. *The Journal of Chemical Physics*, 120(2):516–523, January 2004.

[150] Peter G Bolhuis, David Chandler, Christoph Dellago, and Phillip L Geissler. Transition path sampling: throwing ropes over rough mountain passes, in the dark. *Annual review of physical chemistry*, 53(1):291–318, 2002.

[151] Nicolae-Viorel Buchete and Gerhard Hummer. Coarse master equations for peptide folding dynamics. *The Journal of Physical Chemistry B*, 112(19):6057–6069, May 2008.

[152] Christof Schütte, Frank Noé, Jianfeng Lu, Marco Sarich, and Eric Vanden-Eijnden. Markov state models based on milestoning. *The Journal of Chemical Physics*, 134(20):204105, May 2011.

[153] YounJoon Jung, Juan P Garrahan, and David Chandler. Dynamical exchanges in facilitated models of supercooled liquids. 123(8):084509, August 2005.

[154] Robert E Kass and Adrian E Raftery. Bayes Factors. *Journal of the American Statistical Association*, February 2012.

[155] Ian H Witten and Eibe Frank. Data mining: practical machine learning tools and techniques with java implementations. *Acm Sigmod Record*, 31(1):76–77, 2002.

[156] James S Fraser, Henry van den Bedem, Avi J Samelson, P Therese Lang, James M Holton, Nathaniel Echols, and Tom Alber. Accessing protein conformational ensembles

using room-temperature X-ray crystallography. *Proceedings of the National Academy of Sciences of the United States of America*, 108(39):16247–16252, September 2011.

[157] Jiayin Dai, Shwu-Hwa Lin, Carly Kemmis, Anita J Chin, and J Ching Lee. Interplay between site-specific mutations and cyclic nucleotides in modulating DNA recognition by Escherichia coli cyclic AMP receptor protein. *Biochemistry*, 43(28):8901–8910, July 2004.

[158] H Aiba, T Nakamura, H Mitani, and H Mori. Mutations that alter the allosteric nature of cAMP receptor protein of Escherichia coli. *The EMBO Journal*, 4(12):3329–3332, December 1985.

[159] M Kurplus and J A McCammon. Dynamics of proteins: elements and function. *Annual review of biochemistry*, 1983.

[160] Charles L Brooks, Martin Karplus, and B Montgomery Pettitt. *Advances in chemical physics, volume 71: Proteins: A theoretical perspective of dynamics, structure, and thermodynamics*. Wiley-Blackwell, 2006.

[161] Mark A Depristo, Paul I W de Bakker, and Tom L Blundell. Heterogeneity and Inaccuracy in Protein Structures Solved by X-Ray Crystallography. *Structure*, 12(5):831–838, May 2004.

[162] A J Wand, J L Urbauer, R P McEvoy, and R J Bieber. Internal dynamics of human ubiquitin revealed by 13C-relaxation studies of randomly fractionally labeled protein. *Biochemistry*, 35(19):6116–6125, May 1996.

[163] Kresten Lindorff-Larsen, Robert B Best, Mark A Depristo, Christopher M Dobson, and Michele Vendruscolo. Simultaneous determination of protein structure and dynamics. *Nature*, 433(7022):128–132, January 2005.

[164] Tatyana I Igumenova, Kendra King Frederick, and A Joshua Wand. Characterization of the fast dynamics of protein amino acid side chains using NMR relaxation in solution. *Chemical Reviews*, 106(5):1672–1699, May 2006.

[165] Kateri H Dubay and Phillip L Geissler. Calculation of proteins' total side-chain torsional entropy and its influence on protein-ligand interactions. *Journal of Molecular Biology*, 391(2):484–497, August 2009.

[166] S. Chaudhury, S. Lyskov, and J. J. Gray. PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta. *Bioinformatics*, 26(5):689–691, March 2010.

[167] Maxim V. Shapovalov and Roland L. Dunbrack. A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. *Structure*, 19(6):844–858, June 2011.

[168] W M Oldham and H E Hamm. Structural basis of function in heterotrimeric G proteins. *Quarterly reviews of biophysics*, 2006.

[169] William M Oldham and Heidi E Hamm. Heterotrimeric G protein activation by G-protein-coupled receptors. *Nature reviews. Molecular cell biology*, 9(1):60–71, January 2008.

[170] Christopher A Johnston and David P Siderovski. Receptor-mediated activation of heterotrimeric G-proteins: current structural insights. *Molecular pharmacology*, 72(2):219–230, August 2007.

[171] Tilman Flock, Charles N J Ravarani, Dawei Sun, A J Venkatakrishnan, Melis Kayikci, Christopher G Tate, Dmitry B Veprintsev, and M Madan Babu. Universal allosteric mechanism for G[alpha] activation by GPCRs. *Nature*, 524(7564):173–179, August 2015.

[172] R K Sunahara, J J Tesmer, A G Gilman, and S R Sprang. Crystal structure of the adenylyl cyclase activator Gsalpha. *Science*, 278(5345):1943–1947, December 1997.

[173] Detlef D Leipe, Yuri I Wolf, Eugene V Koonin, and L Aravind. Classification and evolution of P-loop GTPases and related ATPases. *Journal of Molecular Biology*, 317(1):41–72, March 2002.

[174] Gerwin H Westfield, Søren G F Rasmussen, Min Su, Somnath Dutta, Brian T DeVree, Ka Young Chung, Diane Calinski, Gisselle Velez-Ruiz, Austin N Oleskie, Els Pardon, Pil Seok Chae, Tong Liu, Sheng Li, Virgil L Woods, Jan Steyaert, Brian K Kobilka, Roger K Sunahara, and Georgios Skiniotis. Structural flexibility of the G alpha s alpha-helical domain in the beta2-adrenoceptor Gs complex. *Proceedings of the National Academy of Sciences of the United States of America*, 108(38):16086–16091, September 2011.

[175] D G Lambright, J P Noel, H E Hamm, and P B Sigler. Structural determinants for activation of the alpha-subunit of a heterotrimeric G protein. *Nature*, 369(6482):621–628, June 1994.

[176] Yi-Lynn Liang, Maryam Khoshouei, Mazdak Radjainia, Yan Zhang, Alisa Glukhova, Jeffrey Tarrasch, David M Thal, Sebastian G B Furness, George Christopoulos, Thomas Coudrat, Radostin Danev, Wolfgang Baumeister, Laurence J Miller, Arthur Christopoulos, Brian K Kobilka, Denise Wootten, Georgios Skiniotis, and Patrick M Sexton. Phase-plate cryo-EM structure of a class B GPCR–G-protein complex. *Nature*, 159:986, April 2017.

[177] Daniel Hilger, Matthieu Masureel, and Brian K Kobilka. Structure and dynamics of GPCR signaling complexes. *Nature Structural & Molecular Biology*, 25(1):4–12, January 2018.

[178] Sebastian George Barton Furness, Yi-Lynn Liang, Cameron James Nowell, Michelle Louise Halls, Peter John Wookey, Emma Dal Maso, Asuka Inoue, Arthur

Christopoulos, Denise Wootten, and Patrick Michael Sexton. Ligand-Dependent Modulation of G Protein Conformation Alters Drug Efficacy. *Cell*, 167(3):739–749.e11, October 2016.

[179] Yuki Toyama, Hanaho Kano, Yoko Mase, Mariko Yokogawa, Masanori Osawa, and Ichio Shimada. Dynamic regulation of GDP binding to G proteins revealed by magnetic field-dependent NMR relaxation analyses. *Nature communications*, 8:14523, February 2017.

[180] Dawei Sun, Tilman Flock, Xavier Deupi, Shoji Maeda, Milos Matkovic, Sandro Mendieta, Daniel Mayer, Roger J P Dawson, Gebhard F X Schertler, M Madan Babu, and Dmitry B Veprintsev. Probing Gαi1 protein activation at single-amino acid resolution. *Nature Structural & Molecular Biology*, 22(9):686–694, September 2015.

[181] David Goricanec, Ralf Stehle, Pascal Egloff, Simina Grigoriu, Andreas Plückthun, Gerhard Wagner, and Franz Hagn. Conformational dynamics of a G-protein α subunit is tightly regulated by nucleotide binding. *Proceedings of the National Academy of Sciences of the United States of America*, 113(26):E3629–38, June 2016.

[182] Kresten Lindorff-Larsen, Stefano Piana, Ron O Dror, and David E Shaw. How fast-folding proteins fold. *Science*, 334(6055):517–520, October 2011.

[183] Ignasi Buch, Toni Giorgino, and Gianni De Fabritiis. Complete reconstruction of an enzyme-inhibitor binding process by molecular dynamics simulations. *Proceedings of the National Academy of Sciences of the United States of America*, 108(25):10184–10189, June 2011.

[184] Nuria Plattner and Frank Noé. Protein conformational plasticity and complex ligand-binding kinetics explored by atomistic simulations and Markov models. *Nature communications*, 6:7653, July 2015.

[185] Pratyush Tiwary, Vittorio Limongelli, Matteo Salvalaglio, and Michele Parrinello. Kinetics of protein-ligand unbinding: Predicting pathways, rates, and rate-limiting steps.

Proceedings of the National Academy of Sciences of the United States of America, 112(5):E386–91, February 2015.

[186] Nuria Plattner, Stefan Doerr, Gianni De Fabritiis, and Frank Noé. Complete protein–protein association kinetics in atomic detail revealed by molecular dynamics simulations and Markov modelling. *Nature chemistry*, June 2017.

[187] Guangfeng Zhou, George A Pantelopulos, Sudipto Mukherjee, and Vincent A Voelz. Bridging Microscopic and Macroscopic Mechanisms of p53-MDM2 Binding with Kinetic Network Models. *Biophysical Journal*, 113(4):785–793, August 2017.

[188] Ron O Dror, Thomas J Mildorf, Daniel Hilger, Aashish Manglik, David W Borhani, Daniel H Arlow, Ansgar Philippsen, Nicolas Villanueva, Zhongyu Yang, Michael T Lerch, Wayne L Hubbell, Brian K Kobilka, Roger K Sunahara, and David E Shaw. Structural basis for nucleotide exchange in heterotrimeric G proteins. *Science*, 348(6241):1361–1365, June 2015.

[189] Xin-Qiu Yao, Rabia U Malik, Nicholas W Griggs, Lars Skjærven, John R Traynor, Sivaraj Sivaramakrishnan, and Barry J Grant. Dynamic Coupling and Allosteric Networks in the α Subunit of Heterotrimeric G Proteins. *Journal of Biological Chemistry*, 291(9):4742–4753, February 2016.

[190] Peter Chidiac, Vladislav S Markin, and Elliott M Ross. Kinetic control of guanine nucleotide binding to soluble Gαq. *Biochemical pharmacology*, 58(1):39–48, July 1999.

[191] Elliott M Ross. Coordinating speed and amplitude in G-protein signaling. *Current biology : CB*, 18(17):R777–R783, September 2008.

[192] S Mukhopadhyay and E M Ross. Rapid GTP binding and hydrolysis by G(q) promoted by receptor and GTPase-activating proteins. *Proceedings of the National Academy of Sciences*, 96(17):9539–9544, August 1999.

[193] Catherine D Van Raamsdonk, Vladimir Bezrookove, Gary Green, Jürgen Bauer, Lona Gaugler, Joan M O'Brien, Elizabeth M Simpson, Gregory S Barsh, and Boris C Bastian. Frequent somatic mutations of GNAQ in uveal melanoma and blue naevi. *Nature*, 457(7229):599–602, January 2009.

[194] Catherine D Van Raamsdonk, Klaus G Griewank, Michelle B Crosby, Maria C Garrido, Swapna Vemula, Thomas Wiesner, Anna C Obenauf, Werner Wackernagel, Gary Green, Nancy Bouvier, M Mert Sozen, Gail Baimukanova, Ritu Roy, Adriana Heguy, Igor Dolgalev, Raya Khanin, Klaus Busam, Michael R Speicher, Joan O'Brien, and Boris C Bastian. Mutations in GNA11 in uveal melanoma. *The New England journal of medicine*, 363(23):2191–2199, December 2010.

[195] Alessandro Laio and Michele Parrinello. Escaping free-energy minima. *Proceedings of the National Academy of Sciences*, 99(20):12562–12566, October 2002.

[196] Gregory R Bowman, V S Pande, and F Noé. *An introduction to Markov state models and their application to long timescale molecular simulation*. Springer Science & Business Media, 1 edition, 2014.

[197] James S Fraser, Michael W Clarkson, Sheena C Degnan, Renske Erion, Dorothee Kern, and Tom Alber. Hidden alternative structures of proline isomerase essential for catalysis. *Nature*, 462(7273):669–673, December 2009.

[198] Byron Carpenter, Rony Nehmé, Tony Warne, Andrew G W Leslie, and Christopher G Tate. Structure of the adenosine A¡sub¿2A¡/sub¿ receptor bound to an engineered G protein. *Nature*, 536(7614):104–107, August 2016.

[199] D G Lambright, J Sondek, A Bohm, N P Skiba, H E Hamm, and P B Sigler. The 2.0 A crystal structure of a heterotrimeric G protein. *Nature*, 379(6563):311–319, January 1996.

[200] J P Noel, H E Hamm, and P B Sigler. The 2.2 A crystal structure of transducin-alpha complexed with GTP gamma S. *Nature*, 366(6456):654–663, December 1993.

[201] B M Denker, C J Schmidt, and E J Neer. Promotion of the GTP-liganded state of the Go alpha protein by deletion of the C terminus. *Journal of Biological Chemistry*, 267(14):9998–10002, May 1992.

[202] Ethan P Marin, A Gopala Krishna, , and Thomas P Sakmar. Disruption of the α5 Helix of Transducin Impairs Rhodopsin-Catalyzed Nucleotide Exchange†. *Biochemistry*, 41(22):6988–6994, May 2002.

[203] Akiyuki Nishimura, Ken Kitano, Jun Takasaki, Masatoshi Taniguchi, Norikazu Mizuno, Kenji Tago, Toshio Hakoshima, and Hiroshi Itoh. Structural basis for the specific inhibition of heterotrimeric Gq protein by a small molecule. *Proceedings of the National Academy of Sciences of the United States of America*, 107(31):13666–13671, August 2010.

[204] Xuhui Huang, Gregory R Bowman, Sergio Bacallado, and Vijay S Pande. Rapid equilibrium sampling initiated from nonequilibrium data. *Proceedings of the National Academy of Sciences of the United States of America*, 106(47):19765–19769, November 2009.

[205] James F Dama, Michele Parrinello, and Gregory A Voth. Well-tempered metadynamics converges asymptotically. *Physical review letters*, 112(24):240602, June 2014.

[206] Mithun Biswas, Benjamin Lickert, and Gerhard Stock. Metadynamics Enhanced Markov Modeling of Protein Dynamics. *The Journal of Physical Chemistry B*, page acs.jpcb.7b11800, January 2018.

[207] Lu Zhang, Fátima Pardo-Avila, Ilona Christy Unarta, Peter Pak-Hang Cheung, Guo Wang, Dong Wang, and Xuhui Huang. Elucidation of the Dynamics of Transcription Elongation by RNA Polymerase II using Kinetic Network Models. *Accounts of chemical research*, 49(4):687–694, April 2016.

[208] Frank Noé, Christof Schütte, Eric Vanden-Eijnden, Lothar Reich, and Thomas R Weikl. Constructing the equilibrium ensemble of folding pathways from short off-equilibrium

simulations. *Proceedings of the National Academy of Sciences of the United States of America*, 106(45):19011–19016, November 2009.

[209] E Weinan and E Vanden-Eijnden. Towards a Theory of Transition Paths. *Journal of statistical physics*, 2006.

[210] Ned Van Eps, Anita M Preininger, Nathan Alexander, Ali I Kaya, Scott Meier, Jens Meiler, Heidi E Hamm, and Wayne L Hubbell. Interaction of a G protein with an activated receptor opens the interdomain interface in the alpha subunit. *Proceedings of the National Academy of Sciences of the United States of America*, 108(23):9420–9424, June 2011.

[211] Hui-Woog Choe, Yong Ju Kim, Jung Hee Park, Takefumi Morizumi, Emil F Pai, Norbert Krauss, Klaus Peter Hofmann, Patrick Scheerer, and Oliver P Ernst. Crystal structure of metarhodopsin II. *Nature*, 471(7340):651–655, March 2011.

[212] Yan Zhang, Bingfa Sun, Dan Feng, Hongli Hu, Matthew Chu, Qianhui Qu, Jeffrey T Tarrasch, Shane Li, Tong Sun Kobilka, Brian K Kobilka, and Georgios Skiniotis. Cryo-EM structure of the activated GLP-1 receptor in complex with a G protein. *Nature*, 1335:29, May 2017.

[213] William M Oldham, Ned Van Eps, Anita M Preininger, Wayne L Hubbell, and Heidi E Hamm. Mechanism of the receptor-catalyzed activation of heterotrimeric G proteins. *Nature Structural & Molecular Biology*, 13(9):772–777, September 2006.

[214] T Iiri, P Herzmark, J M Nakamoto, C van Dop, and H R Bourne. Rapid GDP release from Gs alpha in patients with gain and loss of endocrine function. *Nature*, 371(6493):164–168, September 1994.

[215] B A Posner, M B Mixon, M A Wall, S R Sprang, and A G Gilman. The A326S mutant of Gialpha1 as an approximation of the receptor-bound state. *Journal of Biological Chemistry*, 273(34):21752–21758, August 1998.

[216] T C Thomas, C J Schmidt, and E J Neer. G-protein alpha o subunit: mutation of conserved cysteines identifies a subunit contact surface and alters GDP affinity. *Proceedings of the National Academy of Sciences*, 90(21):10295–10299, November 1993.

[217] Tilman Flock, Alexander S Hauser, Nadia Lund, David E Gloriam, Santhanam Balaji, and M Madan Babu. Selectivity determinants of GPCR-G-protein binding. *Nature*, 88:263, May 2017.

[218] Mark E Hatley, Steve W Lockless, Scott K Gibson, Alfred G Gilman, and Rama Ranganathan. Allosteric determinants in guanine nucleotide-binding proteins. *Proceedings of the National Academy of Sciences*, 100(24):14445–14450, November 2003.

[219] Rolf Herrmann, Martin Heck, Petra Henklein, Peter Henklein, P Henklein, Christiane Kleuss, Klaus Peter Hofmann, and Oliver P Ernst. Sequence of interactions in receptor-G protein coupling. *Journal of Biological Chemistry*, 279(23):24283–24290, June 2004.

[220] Anita M Preininger, Jens Meiler, and Heidi E Hamm. Conformational flexibility and structural dynamics in GPCR-mediated G protein activation: a perspective. *Journal of Molecular Biology*, 425(13):2288–2298, July 2013.

[221] B Zhang, Y Zhang, Z Wang, and Y Zheng. The role of Mg2+ cofactor in the guanine nucleotide exchange and GTP hydrolysis reactions of Rho family GTP-binding proteins. *Journal of Biological Chemistry*, 275(33):25299–25307, August 2000.

[222] Mark James Abraham, Teemu Murtola, Roland Schulz, Szilárd Páll, Jeremy C Smith, Berk Hess, and Erik Lindahl. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, 1-2:19–25, September 2015.

[223] Kristin L Meagher, Luke T Redman, and Heather A Carlson. Development of polyphosphate parameters for use with the AMBER force field. *Journal of computational chemistry*, 24(9):1016–1025, July 2003.

[224] Giovanni Bussi, Davide Donadio, and Michele Parrinello. Canonical sampling through velocity rescaling. *The Journal of Chemical Physics*, 126(1):014101, January 2007.

[225] Jiri Kolafa and John W Perram. Cutoff Errors in the Ewald Summation Formulae for Point Charge Systems. *Molecular Simulation*, 9(5):351–368, October 2006.

[226] Berk Hess. P-LINCS: A Parallel Linear Constraint Solver for Molecular Simulation. *Journal of Chemical Theory and Computation*, 4(1):116–122, January 2008.

[227] M Parrinello and A Rahman. Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied physics*, 52(12):7182–7190, December 1981.

[228] K A Feenstra, B Hess, and HJC Berendsen. Improving efficiency of large timescale molecular dynamics simulations of hydrogen-rich systems. *J Comput Chem*, 20(8):786–798, June 1999.

[229] Gareth A Tribello, Massimiliano Bonomi, Davide Branduardi, Carlo Camilloni, and Giovanni Bussi. PLUMED 2: New feathers for an old bird. *Computer Physics Communications*, 185(2):604–613, February 2014.

[230] Kyle A Beauchamp, Gregory R Bowman, Thomas J Lane, Lutz Maibaum, Imran S Haque, and Vijay S Pande. MSMBuilder2: Modeling Conformational Dynamics at the Picosecond to Millisecond Scale. *Journal of Chemical Theory and Computation*, 7(10):3412–3419, October 2011.

[231] Luca Maragliano, Alexander Fischer, Eric Vanden-Eijnden, and Giovanni Ciccotti. String method in collective variables: minimum free energy paths and isocommittor surfaces. *The Journal of chemical physics*, 125(2):24106, July 2006.

[232] Luca Maragliano and Eric Vanden-Eijnden. On-the-fly string method for minimum free energy paths calculation. *Chemical physics letters*, 446(1-3):182–190, September 2007.

[233] Luca Maragliano, Benoît Roux, and Eric Vanden-Eijnden. Comparison between Mean Forces and Swarms-of-Trajectories String Methods. *Journal of Chemical Theory and Computation*, 10(2):524–533, February 2014.

[234] Gregory R Bowman, Xuhui Huang, and Vijay S Pande. Using generalized ensemble simulations and Markov state models to identify conformational states. *Methods (San Diego, Calif.)*, 49(2):197–201, October 2009.

[235] Fu Kit Sheong, Daniel-Adriano Silva, Luming Meng, Yutong Zhao, and Xuhui Huang. Automatic state partitioning for multibody systems (APM): an efficient algorithm for constructing Markov state models to elucidate conformational dynamics of multibody systems. *Journal of Chemical Theory and Computation*, 11(1):17–27, January 2015.

[236] William C Swope, Jed W Pitera, and Frank Suits. Describing Protein Folding Kinetics by Molecular Dynamics Simulations. 1. Theory †. *The Journal of Physical Chemistry B*, 108(21):6571–6581, May 2004.

[237] C E Shannon. A Mathematical Theory of Communication. *Bell Labs Technical Journal*, 27(3):379–423, July 1948.

[238] Philipp Metzner, Christof Sch tte, and Eric Vanden-Eijnden. Transition Path Theory for Markov Jump Processes. *Multiscale Modeling & Simulation*, 7(3):1192–1219, January 2009.

[239] Andrew L. Hopkins and Colin R. Groom. The druggable genome. *Nature Reviews Drug Discovery*, 1(9):727–730, September 2002.

[240] David K. Johnson and John Karanicolas. Computational screening and design for compounds that disrupt protein-protein interactions. *Current Topics in Medicinal Chemistry*, 17(23):2703–2714, August 2017.

[241] Sandor Vajda, Dmitri Beglov, Amanda E Wakefield, Megan Egbert, and Adrian Whitty. Cryptic binding sites on proteins: definition, detection, and druggability. *Current Opinion in Chemical Biology*, 44:1–8, June 2018.

[242] D. A. Erlanson, A. C. Braisted, D. R. Raphael, M. Randal, R. M. Stroud, E. M. Gordon, and J. A. Wells. Site-directed ligand discovery. *Proceedings of the National Academy of Sciences*, 97(17):9367–9372, August 2000.

[243] Daniel A Keedy, Zachary B Hill, Justin T Biel, Emily Kang, T Justin Rettenmaier, José Brandão-Neto, Nicholas M Pearce, Frank von Delft, James A Wells, and James S Fraser. An expanded allosteric network in PTP1B by multitemperature crystallography, fragment screening, and covalent tethering. *eLife*, 7:e36307, June 2018.

[244] David K. Johnson and John Karanicolas. Druggable protein interaction sites are more predisposed to surface pocket formation than the rest of the protein surface. *PLoS Computational Biology*, 9(3):e1002951, March 2013.

[245] Vladimiras Oleinikovas, Giorgio Saladino, Benjamin P. Cossins, and Francesco L. Gervasio. Understanding cryptic pocket formation in protein targets by enhanced sampling simulations. *Journal of the American Chemical Society*, 138(43):14257–14263, November 2016.

[246] Peter Cimermancic, Patrick Weinkam, T. Justin Rettenmaier, Leon Bichmann, Daniel A. Keedy, Rahel A. Woldeyes, Dina Schneidman-Duhovny, Omar N. Demerdash, Julie C. Mitchell, James A. Wells, James S. Fraser, and Andrej Sali. Cryptosite: expanding the druggable proteome by characterization and prediction of cryptic binding sites. *Journal of Molecular Biology*, 428(4):709–719, February 2016.

[247] Denis Schmidt, Markus Boehm, Christopher L. McClendon, Rubben Torella, and Holger Gohlke. Cosolvent-enhanced sampling and unbiased identification of cryptic pockets suitable for structure-based drug design. *Journal of Chemical Theory and Computation*, 15(5):3331–3343, May 2019.

[248] Rémi Cuchillo, Kevin Pinto-Gil, and Julien Michel. A collective variable for the rapid exploration of protein druggability. *Journal of Chemical Theory and Computation*, 11(3):1292–1307, March 2015.

[249] Phani Ghanakota and Heather A. Carlson. Moving beyond active-site detection: mixmd applied to allosteric systems. *The Journal of Physical Chemistry B*, 120(33):8685–8695, August 2016.

[250] Christopher D. Wassman, Roberta Baronio, Özlem Demir, Brad D. Wallentine, Chiung-Kuang Chen, Linda V. Hall, Faezeh Salehi, Da-Wei Lin, Benjamin P. Chung, G. Wesley Hatfield, A. Richard Chamberlin, Hartmut Luecke, Richard H. Lathrop, Peter Kaiser, and Rommie E. Amaro. Computational identification of a transiently open L1/S3 pocket for reactivation of mutant p53. *Nature Communications*, 4(1):1407, June 2013.

[251] Rohan Keshwara, Reed F. Johnson, and Matthias J. Schnell. Toward an effective ebola virus vaccine. *Annual Review of Medicine*, 68(1):371–386, January 2017.

[252] Sabue Mulangu, Lori E. Dodd, Richard T. Davey, Olivier Tshiani Mbaya, Michael Proschan, Daniel Mukadi, Mariano Lusakibanza Manzo, Didier Nzolo, Antoine Tshomba Oloma, Augustin Ibanda, Rosine Ali, Sinaré Coulibaly, Adam C. Levine, Rebecca Grais, Janet Diaz, H. Clifford Lane, Jean-Jacques Muyembe-Tamfum, and the PALM Writing Group. A randomized, controlled trial of ebola virus disease therapeutics. *New England Journal of Medicine*, 381(24):2293–2303, December 2019.

[253] Ilhem Messaoudi, Gaya K. Amarasinghe, and Christopher F. Basler. Filovirus pathogenesis and immune evasion: insights from Ebola virus and Marburg virus. *Nature Reviews Microbiology*, 13(11):663–676, November 2015.

[254] Washington B. Cardenas, Yueh-Ming Loo, Michael Gale, Amy L. Hartman, Christopher R. Kimberlin, Luis Martinez-Sobrido, Erica Ollmann Saphire, and Christopher F. Basler. Ebola virus vp35 protein binds double-stranded rna and inhibits alpha/beta in-

terferon production induced by rig-i signaling. *Journal of Virology*, 80(11):5168–5178, June 2006.

[255] Christopher F. Basler, Andrea Mikulasova, Luis Martinez-Sobrido, Jason Paragas, Elke Muhlberger, Mike Bray, Hans-Dieter Klenk, Peter Palese, and Adolfo Garcia-Sastre. The ebola virus vp35 protein inhibits activation of interferon regulatory factor 3. *Journal of Virology*, 77(14):7945–7956, July 2003.

[256] Amy L. Hartman, Jonathan S. Towner, and Stuart T. Nichol. A C-terminal basic amino acid motif of Zaire ebolavirus VP35 is essential for type I interferon antagonism and displays high identity with the RNA-binding domain of another interferon antagonist, the NS1 protein of influenza A virus. *Virology*, 328(2):177–184, October 2004.

[257] D. W. Leung, N. D. Ginder, D. B. Fulton, J. Nix, C. F. Basler, R. B. Honzatko, and G. K. Amarasinghe. Structure of the Ebola VP35 interferon inhibitory domain. *Proceedings of the National Academy of Sciences*, 106(2):411–416, January 2009.

[258] Daisy W Leung, Kathleen C Prins, Dominika M Borek, Mina Farahbakhsh, JoAnn M Tufariello, Parameshwaran Ramanan, Jay C Nix, Luke A Helgeson, Zbyszek Otwinowski, Richard B Honzatko, Christopher F Basler, and Gaya K Amarasinghe. Structural basis for dsRNA recognition and interferon antagonism by Ebola VP35. *Nature Structural & Molecular Biology*, 17(2):165–172, February 2010.

[259] Megan R. Edwards, Gai Liu, Chad E. Mire, Suhas Sureshchandra, Priya Luthra, Benjamin Yen, Reed S. Shabman, Daisy W. Leung, Ilhem Messaoudi, Thomas W. Geisbert, Gaya K. Amarasinghe, and Christopher F. Basler. Differential regulation of interferon responses by ebola and marburg virus vp35 proteins. *Cell Reports*, 14(7):1632–1640, February 2016.

[260] Amy L. Hartman, Jason E. Dover, Jonathan S. Towner, and Stuart T. Nichol. Reverse genetic generation of recombinant zaire ebola viruses containing disrupted irf-3 inhibitory domains results in attenuated virus growth in vitro and higher levels of irf-3 activation

without inhibiting viral transcription or replication. *Journal of Virology*, 80(13):6430–6440, July 2006.

[261] Kathleen C. Prins, Sebastien Delpeut, Daisy W. Leung, Olivier Reynard, Valentina A. Volchkova, St. Patrick Reid, Parameshwaran Ramanan, Washington B. Cardenas, Gaya K. Amarasinghe, Viktor E. Volchkov, and Christopher F. Basler. Mutations abrogating vp35 interaction with double-stranded rna render ebola virus avirulent in guinea pigs. *Journal of Virology*, 84(6):3004–3015, March 2010.

[262] Kathleen C. Prins, Jennifer M. Binning, Reed S. Shabman, Daisy W. Leung, Gaya K. Amarasinghe, and Christopher F. Basler. Basic residues within the ebolavirus vp35 protein are required for its viral polymerase cofactor function. *Journal of Virology*, 84(20):10581–10591, October 2010.

[263] Craig S. Brown, Michael S. Lee, Daisy W. Leung, Tianjiao Wang, Wei Xu, Priya Luthra, Manu Anantpadma, Reed S. Shabman, Lisa M. Melito, Karen S. MacMillan, Dominika M. Borek, Zbyszek Otwinowski, Parameshwaran Ramanan, Alisha J. Stubbs, Dayna S. Peterson, Jennifer M. Binning, Marco Tonelli, Mark A. Olson, Robert A. Davey, Joseph M. Ready, Christopher F. Basler, and Gaya K. Amarasinghe. In silico derived small molecules bind the filovirus vp35 protein and inhibit its polymerase cofactor activity. *Journal of Molecular Biology*, 426(10):2045–2058, May 2014.

[264] Jason G. Glanzer, Brendan M. Byrne, Aaron M. McCoy, Ben J. James, Joshua D. Frank, and Greg G. Oakley. In silico and in vitro methods to identify ebola virus VP35-dsRNA inhibitors. *Bioorganic & Medicinal Chemistry*, 24(21):5388–5392, November 2016.

[265] Vijay S. Pande, Kyle Beauchamp, and Gregory R. Bowman. Everything you wanted to know about Markov State Models but were afraid to ask. *Methods*, 52(1):99–105, September 2010.

[266] Gregory R Bowman, Xuhui Huang, and Vijay S Pande. Network models for molecular kinetics and their initial applications to human health. *Cell Research*, 20(6):622–630, June 2010.

[267] Eugen Hruska, Jayvee R Abella, Feliks Nüske, Lydia E Kavraki, and Cecilia Clementi. Quantitative comparison of adaptive sampling methods for protein dynamics. *The Journal of chemical physics*, 149(24):244119, December 2018.

[268] Dmitri Beglov, David R Hall, Amanda E Wakefield, Lingqi Luo, Karen N Allen, Dima Kozakov, Adrian Whitty, and Sandor Vajda. Exploring the structural origins of cryptic sites on proteins. *Proceedings of the National Academy of Sciences of the United States of America*, 115(15):E3416–E3425, March 2018.

[269] C. H. Ngan, T. Bohnuud, S. E. Mottarella, D. Beglov, E. A. Villar, D. R. Hall, D. Kozakov, and S. Vajda. FTMAP: extended protein mapping with user-selected probe molecules. *Nucleic Acids Research*, 40(W1):W271–W275, July 2012.

[270] Dima Kozakov, Laurie E Grove, David R Hall, Tanggis Bohnuud, Scott E Mottarella, Lingqi Luo, Bing Xia, Dmitri Beglov, and Sandor Vajda. The FTMap family of web servers for determining and characterizing ligand-binding hot spots of proteins. *Nature Protocols*, 10(5):733–755, May 2015.

[271] R. Bernstein, K. L. Schmidt, P. B. Harbury, and S. Marqusee. Structural and kinetic mapping of side-chain exposure onto the protein energy landscape. *Proceedings of the National Academy of Sciences*, 108(26):10532–10537, June 2011.

[272] Gai Liu, Peter J. Nash, Britney Johnson, Colette Pietzsch, Ma. Xenia G. Ilagan, Alexander Bukreyev, Christopher F. Basler, Terry L. Bowlin, Donald T. Moir, Daisy W. Leung, and Gaya K. Amarasinghe. A sensitive in vitro high-throughput screen to identify panfiloviral replication inhibitors targeting the vp35–np interface. *ACS Infectious Diseases*, 3(3):190–198, March 2017.

[273] Alexander G. Kozlov, Roberto Galletto, and Timothy M. Lohman. Ssb–dna binding monitored by fluorescence intensity and anisotropy. In James L. Keck, editor, *Single-Stranded DNA Binding Proteins*, pages 55–83. Humana Press, Totowa, NJ, 2012.

[274] P. Ramanan, M. R. Edwards, R. S. Shabman, D. W. Leung, A. C. Endlich-Frazier, D. M. Borek, Z. Otwinowski, G. Liu, J. Huh, C. F. Basler, and G. K. Amarasinghe. Structural basis for Marburg virus VP35-mediated immune evasion mechanisms. *Proceedings of the National Academy of Sciences*, 109(50):20661–20666, December 2012.

[275] A. Shrake and J.A. Rupley. Environment and exposure to solvent of protein atoms. Lysozyme and insulin. *Journal of Molecular Biology*, 79(2):351–371, September 1973.

[276] Brendan J Frey and Delbert Dueck. Clustering by Passing Messages Between Data Points. *Science*, 315(5814):972–976, February 2007.

[277] Navid Dianati. Unwinding the hairball graph: Pruning algorithms for weighted complex networks. *Physical review. E*, 93(1):012304, January 2016.

[278] Maxwell I. Zimmerman, Kathryn M. Hart, Carrie A. Sibbald, Thomas E. Frederick, John R. Jimah, Catherine R. Knoverek, Niraj H. Tolia, and Gregory R. Bowman. Prediction of new stabilizing mutations based on mechanistic insights from markov state models. *ACS Central Science*, 3(12):1311–1321, December 2017.

[279] Na Zhu, Dingyu Zhang, Wenling Wang, Xingwang Li, Bo Yang, Jingdong Song, Xiang Zhao, Baoying Huang, Weifeng Shi, Roujian Lu, Peihua Niu, Faxian Zhan, Xuejun Ma, Dayan Wang, Wenbo Xu, Guizhen Wu, George F Gao, Wenjie Tan, and China Novel Coronavirus Investigating and Research Team. A novel coronavirus from patients with pneumonia in china, 2019. *N. Engl. J. Med.*, 382(8):727–733, February 2020.

[280] Victor M Corman, Doreen Muth, Daniela Niemeyer, and Christian Drosten. Chapter eight - hosts and sources of endemic human coronaviruses. In Margaret Kielian, Thomas C Mettenleiter, and Marilyn J Roossinck, editors, *Advances in Virus Research*, volume 100, pages 163–188. Academic Press, January 2018.

[281] Max Roser, Hannah Ritchie, Esteban Ortiz-Ospina, and Joe Hasell. Coronavirus pandemic (COVID-19). *Our World in Data*, 2020.

[282] Nicole Lurie, Melanie Saville, Richard Hatchett, and Jane Halton. Developing covid-19 vaccines at pandemic speed. *N. Engl. J. Med.*, 382(21):1969–1973, May 2020.

[283] David E Gordon, Gwendolyn M Jang, Mehdi Bouhaddou, Jiewei Xu, Kirsten Obernier, Kris M White, Matthew J O'Meara, Veronica V Rezelj, Jeffrey Z Guo, Danielle L Swaney, Tia A Tummino, Ruth Huettenhain, Robyn M Kaake, Alicia L Richards, Beril Tutuncuoglu, Helene Foussard, Jyoti Batra, Kelsey Haas, Maya Modak, Minkyu Kim, Paige Haas, Benjamin J Polacco, Hannes Braberg, Jacqueline M Fabius, Manon Eckhardt, Margaret Soucheray, Melanie J Bennett, Merve Cakir, Michael J McGregor, Qiongyu Li, Bjoern Meyer, Ferdinand Roesch, Thomas Vallet, Alice Mac Kain, Lisa Miorin, Elena Moreno, Zun Zar Chi Naing, Yuan Zhou, Shiming Peng, Ying Shi, Ziyang Zhang, Wenqi Shen, Ilsa T Kirby, James E Melnyk, John S Chorba, Kevin Lou, Shizhong A Dai, Inigo Barrio-Hernandez, Danish Memon, Claudia Hernandez-Armenta, Jiankun Lyu, Christopher J P Mathy, Tina Perica, Kala B Pilla, Sai J Ganesan, Daniel J Saltzberg, Ramachandran Rakesh, Xi Liu, Sara B Rosenthal, Lorenzo Calviello, Srivats Venkataramanan, Jose Liboy-Lugo, Yizhu Lin, Xi-Ping Huang, Yongfeng Liu, Stephanie A Wankowicz, Markus Bohn, Maliheh Safari, Fatima S Ugur, Cassandra Koh, Nastaran Sadat Savar, Quang Dinh Tran, Djoshkun Shengjuler, Sabrina J Fletcher, Michael C O'Neal, Yiming Cai, Jason C J Chang, David J Broadhurst, Saker Klippsten, Phillip P Sharp, Nicole A Wenzell, Duygu Kuzuoglu, Hao-Yuan Wang, Raphael Trenker, Janet M Young, Devin A Cavero, Joseph Hiatt, Theodore L Roth, Ujjwal Rathore, Advait Subramanian, Julia Noack, Mathieu Hubert, Robert M Stroud, Alan D Frankel, Oren S Rosenberg, Kliment A Verba, David A Agard, Melanie Ott, Michael Emerman, Natalia Jura, Mark von Zastrow, Eric Verdin, Alan Ashworth, Olivier Schwartz, Christophe d'Enfert, Shaeri Mukherjee, Matt Jacobson, Harmit S Malik, Danica G Fujimori, Trey Ideker, Charles S Craik, Stephen N Floor, James S Fraser, John D Gross, Andrej Sali,

Bryan L Roth, Davide Ruggero, Jack Taunton, Tanja Kortemme, Pedro Beltrao, Marco Vignuzzi, Adolfo García-Sastre, Kevan M Shokat, Brian K Shoichet, and Nevan J Krogan. A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature*, April 2020.

[284] James M Sanders, Marguerite L Monogue, Tomasz Z Jodlowski, and James B Cutrell. Pharmacologic treatments for coronavirus disease 2019 (COVID-19): A review. *JAMA*, April 2020.

[285] Alexandra C Walls, Young-Jun Park, M Alejandra Tortorici, Abigail Wall, Andrew T McGuire, and David Veesler. Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. *Cell*, 181(2):281–292.e6, April 2020.

[286] Markus Hoffmann, Hannah Kleine-Weber, Simon Schroeder, Nadine Krüger, Tanja Herrler, Sandra Erichsen, Tobias S Schiergens, Georg Herrler, Nai-Huei Wu, Andreas Nitsche, Marcel A Müller, Christian Drosten, and Stefan Pöhlmann. SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor. *Cell*, 181(2):271–280.e8, April 2020.

[287] Jian Shang, Gang Ye, Ke Shi, Yushun Wan, Chuming Luo, Hideki Aihara, Qibin Geng, Ashley Auerbach, and Fang Li. Structural basis of receptor recognition by SARS-CoV-2. *Nature*, 581(7807):221–224, May 2020.

[288] Jun Lan, Jiwan Ge, Jinfang Yu, Sisi Shan, Huan Zhou, Shilong Fan, Qi Zhang, Xuanling Shi, Qisheng Wang, Linqi Zhang, and Xinquan Wang. Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. *Nature*, 581(7807):215–220, May 2020.

[289] Paul S. Masters. Coronavirus genomic RNA packaging. *Virology*, 537:198–207, November 2019.

[290] Robin van der Lee, Marija Buljan, Benjamin Lang, Robert J Weatheritt, Gary W Daughdrill, A Keith Dunker, Monika Fuxreiter, Julian Gough, Joerg Gsponer, David T Jones,

Philip M Kim, Richard W Kriwacki, Christopher J Oldfield, Rohit V Pappu, Peter Tompa, Vladimir N Uversky, Peter E Wright, and M Madan Babu. Classification of intrinsically disordered regions and proteins. *Chem. Rev.*, 114(13):6589–6631, July 2014.

[291] Hubert Laude and Paul S Masters. The coronavirus nucleocapsid protein. In Stuart G Siddell, editor, *The Coronaviridae*, pages 141–163. Springer US, Boston, MA, 1995.

[292] Erik D Holmstrom, Zhaowei Liu, Daniel Nettels, Robert B Best, and Benjamin Schuler. Disordered RNA chaperones can enhance nucleic acid folding via local charge screening. *Nat. Commun.*, 10(1):2453, June 2019.

[293] Alessandro Borgia, Madeleine B Borgia, Katrine Bugge, Vera M Kissling, Pétur O Heidarsson, Catarina B Fernandes, Andrea Sottini, Andrea Soranno, Karin J Buholzer, Daniel Nettels, Birthe B Kragelund, Robert B Best, and Benjamin Schuler. Extreme disorder in an ultrahigh-affinity protein complex. *Nature*, 555:61, February 2018.

[294] Mykola Dimura, Thomas O Peulen, Christian A Hanke, Aiswaria Prakash, Holger Gohlke, and Claus Am Seidel. Quantitative FRET studies and integrative modeling unravel the structure and dynamics of biomolecular systems. *Curr. Opin. Struct. Biol.*, 40:163–185, October 2016.

[295] Gustavo Fuertes, Niccolò Banterle, Kiersten M Ruff, Aritra Chowdhury, Davide Mercadante, Christine Koehler, Michael Kachala, Gemma Estrada Girona, Sigrid Milles, Ankur Mishra, Patrick R Onck, Frauke Gräter, Santiago Esteban-Martín, Rohit V Pappu, Dmitri I Svergun, and Edward A Lemke. Decoupling of size and shape fluctuations in heteropolymeric sequences reconciles discrepancies in SAXS vs. FRET measurements. *Proc. Natl. Acad. Sci. U. S. A.*, 114(31):E6342–E6351, August 2017.

[296] John B Warner, 4th, Kiersten M Ruff, Piau Siong Tan, Edward A Lemke, Rohit V Pappu, and Hilal A Lashuel. Monomeric huntingtin exon 1 has similar overall structural features for Wild-Type and pathological polyglutamine lengths. *J. Am. Chem. Soc.*, 139(41):14456–14469, October 2017.

[297] Hoi Sung Chung, Stefano Piana-Agostinetti, David E Shaw, and William A Eaton. Structural origin of slow diffusion in protein folding. *Science*, 349(6255):1504–1510, September 2015.

[298] Christiane Iserman, Christine Anne Roden, Mark Boerneke, Rachel Sealfon, Grace McLaughlin, Irwin Jungreis, Christopher Y Park, Avinash Boppana, Ethan Fritch, Yixuan Hou, Chandra Theesfeld, Olga Troyanskaya, Ralph S G Baric, Timothy P Sheahan, Kevin Weeks, and Amy Gladfelter. Specific viral RNA drives the SARS CoV-2 nucleocapsid to phase separate. June 2020.

[299] Theodora Myrto Perdikari, Anastasia C Murthy, Veronica H Ryan, Scott Watters, Mandar T Naik, and Nicolas L Fawzi. SARS-CoV-2 nucleocapsid protein undergoes liquid-liquid phase separation stimulated by RNA and partitions into phases of human ribonucleoproteins. June 2020.

[300] Adriana Savastano, Alain Ibáñez de Opakua, Marija Rankovic, and Markus Zweckstetter. Nucleocapsid protein of SARS-CoV-2 phase separates into RNA-rich polymerase-containing condensates. June 2020.

[301] Ruth McBride, Marjorie van Zyl, and Burtram Fielding. The coronavirus nucleocapsid is a multifunctional protein. *Viruses*, 6(8):2991–3018, August 2014.

[302] N A Baker, D Sept, S Joseph, M J Holst, and J A McCammon. Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc. Natl. Acad. Sci. U. S. A.*, 98(18):10037–10041, August 2001.

[303] Chung-Ke Chang, Yen-Lan Hsu, Yuan-Hsiang Chang, Fa-An Chao, Ming-Chya Wu, Yu-Shan Huang, Chin-Kun Hu, and Tai-Huang Huang. Multiple nucleic acid binding sites and intrinsic disorder of severe acute respiratory syndrome coronavirus nucleocapsid protein: implications for ribonucleocapsid protein packaging. *J. Virol.*, 83(5):2255–2264, March 2009.

[304] Nicholas E Grossoehme, Lichun Li, Sarah C Keane, Pinghua Liu, Charles E Dann, 3rd, Julian L Leibowitz, and David P Giedroc. Coronavirus N protein n-terminal domain (NTD) specifically binds the transcriptional regulatory sequence (TRS) and melts TRS-cTRS RNA duplexes. *J. Mol. Biol.*, 394(3):544–557, December 2009.

[305] Lei Cui, Haiying Wang, Yanxi Ji, Jie Yang, Shan Xu, Xingyu Huang, Zidao Wang, Lei Qin, Po Tien, Xi Zhou, Deyin Guo, and Yu Chen. The nucleocapsid protein of coronaviruses acts as a viral suppressor of RNA silencing in mammalian cells. *J. Virol.*, 89(17):9029–9043, September 2015.

[306] Mitsuhiro Takeda, Chung-Ke Chang, Teppei Ikeya, Peter Güntert, Yuan-Hsiang Chang, Yen-Lan Hsu, Tai-Huang Huang, and Masatsune Kainosho. Solution structure of the c-terminal dimerization domain of SARS coronavirus nucleocapsid protein solved by the SAIL-NMR method. *J. Mol. Biol.*, 380(4):608–622, July 2008.

[307] Hariharan Jayaram, Hui Fan, Brian R Bowman, Amy Ooi, Jyothi Jayaram, Ellen W Collisson, Julien Lescar, and B V Venkataram Prasad. X-ray structures of the N- and c-terminal domains of a coronavirus nucleocapsid protein: implications for nucleocapsid formation. *J. Virol.*, 80(13):6612–6620, July 2006.

[308] I-Mei Yu, Christin L T Gustafson, Jianbo Diao, John W Burgner, 2nd, Zhihong Li, Jingqiang Zhang, and Jue Chen. Recombinant severe acute respiratory syndrome (SARS) coronavirus nucleocapsid protein forms a dimer through its c-terminal domain. *J. Biol. Chem.*, 280(24):23280–23286, June 2005.

[309] Haibin Luo, Jing Chen, Kaixian Chen, Xu Shen, and Hualiang Jiang. Carboxyl terminus of severe acute respiratory syndrome coronavirus nucleocapsid protein: self-association analysis and nucleic acid binding characterization. *Biochemistry*, 45(39):11827–11835, October 2006.

[310] Chung-Ke Chang, Chia-Min Michael Chen, Ming-Hui Chiang, Yen-Lan Hsu, and Tai-Huang Huang. Transient oligomerization of the SARS-CoV N protein–implication for virus ribonucleoprotein packaging. *PLoS One*, 8(5):e65045, May 2013.

[311] S G Robbins, M F Frana, J J McGowan, J F Boyle, and K V Holmes. RNA-binding proteins of coronavirus MHV: detection of monomeric and multimeric N protein with an RNA overlay-protein blot assay. *Virology*, 150(2):402–410, April 1986.

[312] Runtao He, Frederick Dobie, Melissa Ballantine, Andrew Leeson, Yan Li, Nathalie Bastien, Todd Cutts, Anton Andonov, Jingxin Cao, Timothy F Booth, Frank A Plummer, Shaun Tyler, Lindsay Baker, and Xuguang Li. Analysis of multimerization of the SARS coronavirus nucleocapsid protein. *Biochem. Biophys. Res. Commun.*, 316(2):476–483, April 2004.

[313] Sisi Kang, Mei Yang, Zhongsi Hong, Liping Zhang, Zhaoxia Huang, Xiaoxue Chen, Suhua He, Ziliang Zhou, Zhechong Zhou, Qiuyue Chen, Yan Yan, Changsheng Zhang, Hong Shan, and Shoudeng Chen. Crystal structure of SARS-CoV-2 nucleocapsid protein RNA binding domain reveals potential unique drug targeting sites. *Acta Pharm Sin B*, April 2020.

[314] Luca Zinzula, Massimiliano Orsini Nagy, and Andreas Bracher. 1.45 angstrom resolution crystal structure of c-terminal dimerization domain of nucleocapsid phosphoprotein from SARS-CoV-2 (PDB: 6YUN). *Protein Data Bank*, May 2020.

[315] Qiaozhen Ye, Alan M V West, Steve Silletti, and Kevin D Corbett. Architecture and self-assembly of the SARS-CoV-2 nucleocapsid protein. June 2020.

[316] Weihong Zeng, Guangfeng Liu, Huan Ma, Dan Zhao, Yunru Yang, Muziying Liu, Ahmed Mohammed, Changcheng Zhao, Yun Yang, Jiajia Xie, Chengchao Ding, Xiaoling Ma, Jianping Weng, Yong Gao, Hongliang He, and Tengchuan Jin. Biochemical characterization of SARS-CoV-2 nucleocapsid protein. *Biochem. Biophys. Res. Commun.*, 527(3):618–623, June 2020.

[317] Daniel Nettels, Sonja Müller-Späth, Frank Küster, Hagen Hofmann, Dominik Haenni, Stefan Rüegger, Luc Reymond, Armin Hoffmann, Jan Kubelka, Benjamin Heinz, Klaus Gast, Robert B Best, and Benjamin Schuler. Single-molecule spectroscopy of the temperature-induced collapse of unfolded proteins. *Proc. Natl. Acad. Sci. U. S. A.*, 106(49):20740–20745, December 2009.

[318] Andrea Soranno, Brigitte Buchli, Daniel Nettels, Ryan R Cheng, Sonja Müller-Späth, Shawn H Pfeil, Armin Hoffmann, Everett A Lipman, Dmitrii E Makarov, and Benjamin Schuler. Quantifying internal friction in unfolded and intrinsically disordered proteins with single-molecule spectroscopy. *Proc. Natl. Acad. Sci. U. S. A.*, 109(44):17800–17806, October 2012.

[319] Alessandro Borgia, Beth G Wensley, Andrea Soranno, Daniel Nettels, Madeleine B Borgia, Armin Hoffmann, Shawn H Pfeil, Everett A Lipman, Jane Clarke, and Benjamin Schuler. Localizing internal friction along the reaction coordinate of protein folding by combining ensemble and single-molecule fluorescence spectroscopy. *Nat. Commun.*, 3:1195, 2012.

[320] Benjamin Schuler, Andrea Soranno, Hagen Hofmann, and Daniel Nettels. Single-Molecule FRET spectroscopy and the polymer physics of unfolded and intrinsically disordered proteins. *Annu. Rev. Biophys.*, 45:207–231, July 2016.

[321] A Soranno, F Cabassi, M E Orselli, and others. Dynamics of structural elements of GB1 β-Hairpin revealed by Tryptophan–Cysteine contact formation experiments. *The Journal of*, 2018.

[322] Andrea Soranno, Andrea Holla, Fabian Dingfelder, Daniel Nettels, Dmitrii E Makarov, and Benjamin Schuler. Integrated view of internal friction in unfolded proteins from single-molecule FRET, contact quenching, theory, and simulations. *Proc. Natl. Acad. Sci. U. S. A.*, 114(10):E1833–E1839, March 2017.

[323] J A Schellman. Selective binding and solvent denaturation. *Biopolymers*, 26(4):549–559, April 1987.

[324] Hagen Hofmann, Andrea Soranno, Alessandro Borgia, Klaus Gast, Daniel Nettels, and Benjamin Schuler. Polymer scaling laws of unfolded and intrinsically disordered proteins quantified with single-molecule spectroscopy. *Proc. Natl. Acad. Sci. U. S. A.*, 109(40):16155–16160, September 2012.

[325] Alessandro Borgia, Wenwei Zheng, Karin Buholzer, Madeleine B Borgia, Anja Schüler, Hagen Hofmann, Andrea Soranno, Daniel Nettels, Klaus Gast, Alexander Grishaev, and Others. Consistent view of polypeptide chain expansion in chemical denaturants from multiple experimental methods. *J. Am. Chem. Soc.*, 138(36):11714–11726, 2016.

[326] Wenwei Zheng, Alessandro Borgia, Karin Buholzer, Alexander Grishaev, Benjamin Schuler, and Robert B Best. Probing the action of chemical denaturant on an intrinsically disordered protein by simulation and experiment. *J. Am. Chem. Soc.*, 138(36):11702–11713, September 2016.

[327] Mikayel Aznauryan, Leonildo Delgado, Andrea Soranno, Daniel Nettels, Jie-Rong Huang, Alexander M Labhardt, Stephan Grzesiek, and Benjamin Schuler. Comprehensive structural and dynamical view of an unfolded protein from the combination of single-molecule FRET, NMR, and SAXS. *Proc. Natl. Acad. Sci. U. S. A.*, 113(37):E5389–98, September 2016.

[328] Peter Tompa and Monika Fuxreiter. Fuzzy complexes: polymorphism and structural disorder in protein–protein interactions. *Trends Biochem. Sci.*, 33(1):2–8, 2008.

[329] Alex S Holehouse, Kanchan Garai, Nicholas Lyle, Andreas Vitalis, and Rohit V Pappu. Quantitative assessments of the distinct contributions of polypeptide backbone amides versus side chain groups to chain expansion via chemical denaturation. *J. Am. Chem. Soc.*, 137(8):2984–2995, March 2015.

[330] Daniel Nettels, Irina V Gopich, Armin Hoffmann, and Benjamin Schuler. Ultrafast dynamics of protein collapse from single-molecule photon statistics. *Proc. Natl. Acad. Sci. U. S. A.*, 104(8):2655–2660, February 2007.

[331] Markus Sauer and Hannes Neuweiler. PET-FCS: probing rapid structural fluctuations of proteins and nucleic acids by single-molecule fluorescence quenching. *Methods Mol. Biol.*, 1076:597–615, 2014.

[332] Dominik Haenni, Franziska Zosel, Luc Reymond, Daniel Nettels, and Benjamin Schuler. Intramolecular distances and dynamics from the combined photon statistics of single-molecule FRET and photoinduced electron transfer. *J. Phys. Chem. B*, 117(42):13015–13028, October 2013.

[333] Franziska Zosel, Dominik Haenni, Andrea Soranno, Daniel Nettels, and Benjamin Schuler. Combining short- and long-range fluorescence reporters with simulations to explore the intramolecular dynamics of an intrinsically disordered protein. *J. Chem. Phys.*, 147(15):152708, October 2017.

[334] Salman F Banani, Hyun O Lee, Anthony A Hyman, and Michael K Rosen. Biomolecular condensates: organizers of cellular biochemistry. *Nat. Rev. Mol. Cell Biol.*, 18(5):285–298, May 2017.

[335] Yongdae Shin and Clifford P Brangwynne. Liquid phase condensation in cell physiology and disease. *Science*, 357(6357), September 2017.

[336] Clifford P Brangwynne, Christian R Eckmann, David S Courson, Agata Rybarska, Carsten Hoege, Joebin Gharakhani, Frank Juelicher, and Anthony A Hyman. Germline P granules are liquid droplets that localize by controlled dissolution/condensation. *Science*, 324(5935):1729–1732, June 2009.

[337] Pilong Li, Sudeep Banjade, Hui-Chun Cheng, Soyeon Kim, Baoyu Chen, Liang Guo, Marc Llaguno, Javoris V Hollingsworth, David S King, Salman F Banani, Paul S Russo,

Qiu-Xing Jiang, B Tracy Nixon, and Michael K Rosen. Phase transitions in the assembly of multivalent signalling proteins. *Nature*, 483(7389):336–340, March 2012.

[338] Erik W Martin, Alex S Holehouse, Ivan Peran, Mina Farag, J Jeremias Incicco, Anne Bremer, Christy R Grace, Andrea Soranno, Rohit V Pappu, and Tanja Mittag. Valence and patterning of aromatic residues determine the phase behavior of prion-like domains. *Science*, 367(6478):694–699, February 2020.

[339] Jordina Guillén-Boixet, Andrii Kopach, Alex S Holehouse, Sina Wittmann, Marcus Jahnel, Raimund Schlüßler, Kyoohyun Kim, Irmela R E, Jie Wang, Daniel Mateju, Ina Poser, Shovamayee Maharana, Martine Ruer-Gruß, Doris Richter, Xiaojie Zhang, Young-Tae Chang, Jochen Guck, Alf Honigmann, Julia Mahamid, Anthony A Hyman, Rohit V Pappu, Simon Alberti, and Titus M Franzmann. RNA-Induced conformational switching and clustering of G3BP drive stress granule assembly by condensation. *Cell*, 181(2):346–361.e17, April 2020.

[340] Jie Wang, Jeong-Mo Choi, Alex S Holehouse, Hyun O Lee, Xiaojie Zhang, Marcus Jahnel, Shovamayee Maharana, Régis Lemaitre, Andrei Pozniakovsky, David Drechsel, Ina Poser, Rohit V Pappu, Simon Alberti, and Anthony A Hyman. A Molecular Grammar Governing the Driving Forces for Phase Separation of Prion-like RNA Binding Proteins. *Cell*, June 2018.

[341] W H Stockmayer. Light scattering in Multi-Component systems. *J. Chem. Phys.*, 18(1):58–61, January 1950.

[342] Priya R Banerjee, Anthony N Milin, Mahdi Muhammad Moosa, Paulo L Onuchic, and Ashok A Deniz. Reentrant phase transition drives dynamic substructure formation in ribonucleoprotein droplets. *Angew. Chem. Int. Ed Engl.*, 56(38):11354–11359, September 2017.

[343] Serafima Guseva, Sigrid Milles, Malene Ringkjøbing Jensen, Nicola Salvi, Jean-Philippe Kleman, Damien Maurin, Rob W H Ruigrok, and Martin Blackledge. Measles

virus nucleo- and phosphoproteins form liquid-like phase-separated compartments that promote nucleocapsid assembly. *Sci Adv*, 6(14):eaaz7095, April 2020.

[344] Ammon E Posey, Alex S Holehouse, and Rohit V Pappu. Chapter one - phase separation of intrinsically disordered proteins. In Elizabeth Rhoades, editor, *Methods in Enzymology*, volume 611, pages 1–30. Academic Press, January 2018.

[345] David W Sanders, Nancy Kedersha, Daniel S W Lee, Amy R Strom, Victoria Drake, Joshua A Riback, Dan Bracha, Jorine M Eeftens, Allana Iwanicki, Alicia Wang, Ming-Tzo Wei, Gena Whitney, Shawn M Lyons, Paul Anderson, William M Jacobs, Pavel Ivanov, and Clifford P Brangwynne. Competing Protein-RNA interaction networks control multiphase intracellular organization. *Cell*, 181(2):306–324.e28, April 2020.

[346] Joshua A Riback, Lian Zhu, Mylene C Ferrolino, Michele Tolbert, Diana M Mitrea, David W Sanders, Ming-Tzo Wei, Richard W Kriwacki, and Clifford P Brangwynne. Composition-dependent thermodynamics of intracellular phase separation. *Nature*, 581(7807):209–214, May 2020.

[347] Alexander N Semenov and Michael Rubinstein. Thermoreversible gelation in solutions of associative polymers. 1. statics. *Macromolecules*, 31(4):1373–1385, February 1998.

[348] M Rubinstein and Ralph H Colby. *Polymer Physics*. Oxford University Press, New York, 2003.

[349] Jeong-Mo Choi, Alex S Holehouse, and Rohit V Pappu. Physical principles underlying the complex biology of intracellular phase transitions. *Annu. Rev. Biophys.*, 49:107–133, May 2020.

[350] Jeong-Mo Choi, Furqan Dar, and Rohit V Pappu. LASSI: A lattice model for simulating phase transitions of multivalent proteins. *PLoS Comput. Biol.*, 15(10):e1007028, October 2019.

[351] Carol Beth Post and Bruno H Zimm. Internal condensation of a single DNA molecule. *Biopolymers*, 18(6):1487–1501, June 1979.

[352] Ping-Kun Hsieh, Shin C Chang, Chu-Chun Huang, Ting-Ting Lee, Ching-Wen Hsiao, Yi-Hen Kou, I-Yin Chen, Chung-Ke Chang, Tai-Huang Huang, and Ming-Fu Chang. Assembly of severe acute respiratory syndrome coronavirus RNA packaging signal into virus-like particles is nucleocapsid dependent. *J. Virol.*, 79(22):13848–13855, November 2005.

[353] K Woo, M Joo, K Narayanan, K H Kim, and S Makino. Murine coronavirus packaging signal confers packaging to nonviral RNA. *J. Virol.*, 71(1):824–827, January 1997.

[354] R Cologna and B G Hogue. Identification of a bovine coronavirus packaging signal. *J. Virol.*, 74(1):580–583, January 2000.

[355] René Pool and Peter G Bolhuis. Sampling the kinetic pathways of a micelle fusion and fission transition. *J. Chem. Phys.*, 126(24):244703, June 2007.

[356] Antonia G Denkova, Eduardo Mendes, and Marc-Olivier Coppens. Non-equilibrium dynamics of block copolymer micelles in solution: recent insights and open questions. *Soft Matter*, 6(11):2351–2357, 2010.

[357] Srivastav Ranganathan and Eugene I Shakhnovich. Dynamic metastable long-living droplets formed by sticker-spacer proteins. *Elife*, 9, June 2020.

[358] Cédric Leyrat, Malene Ringkjøbing Jensen, Euripedes A Ribeiro, Jr, Francine C A Gérard, Rob W H Ruigrok, Martin Blackledge, and Marc Jamin. The n0-binding region of the vesicular stomatitis virus phosphoprotein is globally disordered but contains transient α-helices. *Protein Sci.*, 20(3):542–556, 2011.

[359] Sophie Feuerstein, Zsofia Solyom, Amine Aladag, Adrien Favier, Melanie Schwarten, Silke Hoffmann, Dieter Willbold, and Bernhard Brutscher. Transient structure and SH3

interaction sites in an intrinsically disordered fragment of the hepatitis C virus protein NS5A. *J. Mol. Biol.*, 420(4-5):310–323, July 2012.

[360] Malene Ringkjøbing Jensen, Klaartje Houben, Ewen Lescop, Laurence Blanchard, Rob W H Ruigrok, and Martin Blackledge. Quantitative conformational analysis of partially folded proteins from residual dipolar couplings: application to the molecular recognition element of sendai virus nucleoprotein. *J. Am. Chem. Soc.*, 130(25):8055–8061, 2008.

[361] Travis S Bayer, Lauren N Booth, Scott M Knudsen, and Andrew D Ellington. Arginine-rich motifs present multiple interfaces for specific binding by RNA. *RNA*, 11(12):1848–1857, December 2005.

[362] J L Battiste, H Mao, N S Rao, R Tan, D R Muhandiram, L E Kay, A D Frankel, and J R Williamson. Alpha helix-RNA major groove recognition in an HIV-1 rev peptide-RRE RNA complex. *Science*, 273(5281):1547–1551, September 1996.

[363] Kelley R Hurst, Cheri A Koetzner, and Paul S Masters. Characterization of a critical interaction between the coronavirus nucleocapsid protein and nonstructural protein 3 of the viral replicase-transcriptase complex. *J. Virol.*, 87(16):9159–9172, August 2013.

[364] Kelley R Hurst, Rong Ye, Scott J Goebel, Priya Jayaraman, and Paul S Masters. An interaction between the nucleocapsid protein and a component of the Replicase-Transcriptase complex is crucial for the infectivity of coronavirus genomic RNA. *J. Virol.*, 84(19):10276–10288, October 2010.

[365] Monique H Verheije, Marne C Hagemeijer, Mustafa Ulasli, Fulvio Reggiori, Peter J M Rottier, Paul S Masters, and Cornelis A M de Haan. The coronavirus nucleocapsid protein is dynamically associated with the replication-transcription complexes. *J. Virol.*, 84(21):11575–11579, November 2010.

[366] Milan Surjit, Ravinder Kumar, Rabi N Mishra, Malireddy K Reddy, Vincent T K Chow, and Sunil K Lal. The severe acute respiratory syndrome coronavirus nucleocapsid pro-

tein is phosphorylated and localizes in the cytoplasm by 14-3-3-mediated translocation. *J. Virol.*, 79(17):11476–11486, September 2005.

[367] Khalid Amine Timani, Qingjiao Liao, Linbai Ye, Yingchun Zeng, Jing Liu, Yi Zheng, Li Ye, Xiaojun Yang, Kong Lingbao, Jingrong Gao, and Ying Zhu. Nuclear/nucleolar localization properties of c-terminal nucleocapsid protein of SARS coronavirus. *Virus Res.*, 114(1-2):23–34, December 2005.

[368] Lili Kuo and Paul S Masters. Genetic evidence for a structural interaction between the carboxy termini of the membrane and nucleocapsid proteins of mouse hepatitis virus. *J. Virol.*, 76(10):4987–4999, May 2002.

[369] Kelley R Hurst, Lili Kuo, Cheri A Koetzner, Rong Ye, Bilan Hsue, and Paul S Masters. A major determinant for membrane protein interaction localizes to the carboxy-terminal domain of the mouse coronavirus nucleocapsid protein. *J. Virol.*, 79(21):13285–13297, 2005.

[370] Sandhya Verma, Valerie Bednar, Andrew Blount, and Brenda G Hogue. Identification of functionally important negatively charged residues in the carboxy end of mouse hepatitis coronavirus A59 nucleocapsid protein. *J. Virol.*, 80(9):4344–4355, May 2006.

[371] Volker Brass, Elke Bieck, Roland Montserret, Benno Wölk, Jan Albert Hellings, Hubert E Blum, François Penin, and Darius Moradpour. An amino-terminal amphipathic α-Helix mediates membrane association of the hepatitis C virus nonstructural protein 5A. *J. Biol. Chem.*, 277(10):8130–8139, March 2002.

[372] Anthony R Braun, Michael M Lacy, Vanessa C Ducas, Elizabeth Rhoades, and Jonathan N Sachs. α-Synuclein's uniquely long amphipathic helix enhances its membrane binding and remodeling capacity. *J. Membr. Biol.*, 250(2):183–193, April 2017.

[373] Jeffries Wyman and Stanley J Gill. *Binding and Linkage: Functional Chemistry of Biological Macromolecules*. University Science Books, 1990.

[374] Jovan Nikolic, Romain Le Bars, Zoé Lama, Nathalie Scrima, Cécile Lagaudrière-Gesbert, Yves Gaudin, and Danielle Blondel. Negri bodies are viral factories with properties of liquid organelles. *Nat. Commun.*, 8(1):58, July 2017.

[375] Claire M Metrick, Andrea L Koenigsberg, and Ekaterina E Heldwein. Conserved outer tegument component UL11 from herpes simplex virus 1 is an intrinsically disordered, RNA-Binding protein. *MBio*, 11(3), May 2020.

[376] Bianca S Heinrich, Zoltan Maliga, David A Stein, Anthony A Hyman, and Sean P J Whelan. Phase transitions drive the formation of vesicular stomatitis virus replication compartments. *MBio*, 9(5), September 2018.

[377] Yuqin Zhou, Justin M Su, Charles E Samuel, and Dzwokai Ma. Measles virus forms inclusion bodies with properties of liquid organelles. *J. Virol.*, 93(21), November 2019.

[378] Anne Monette, Meijuan Niu, Lois Chen, Shringar Rao, Robert James Gorelick, and Andrew John Mouland. Pan-retroviral Nucleocapsid-Mediated phase separation regulates genomic RNA positioning and trafficking. *Cell Rep.*, 31(3):107520, April 2020.

[379] Steffen Klein, Mirko Cortese, Sophie L Winter, Moritz Wachsmuth-Melm, Christopher J Neufeldt, Berati Cerikan, Megan L Stanifer, Steeve Boulant, Ralf Bartenschlager, and Petr Chlanda. SARS-CoV-2 structure and replication characterized by in situ cryo-electron tomography. June 2020.

[380] Yingying Cong, Franziska Kriegenburg, Cornelis A M de Haan, and Fulvio Reggiori. Coronavirus nucleocapsid proteins assemble constitutively in high molecular oligomers. *Sci. Rep.*, 7(1):5740, July 2017.

[381] Chung-Ke Chang, Ming-Hon Hou, Chi-Fon Chang, Chwan-Deng Hsiao, and Tai-Huang Huang. The SARS coronavirus nucleocapsid protein–forms and functions. *Antiviral Res.*, 103:39–50, 2014.

[382] Alexander Borodavka, Roman Tuma, and Peter G Stockley. A two-stage mechanism of viral RNA compaction revealed by single molecule fluorescence. *RNA Biol.*, 10(4):481–489, April 2013.

[383] Runtao He, Andrew Leeson, Melissa Ballantine, Anton Andonov, Lindsay Baker, Frederick Dobie, Yan Li, Nathalie Bastien, Heinz Feldmann, Ute Strocher, Steven Theriault, Todd Cutts, Jingxin Cao, Timothy F Booth, Frank A Plummer, Shaun Tyler, and Xuguang Li. Characterization of protein–protein interactions between the nucleocapsid protein and membrane protein of the SARS coronavirus. *Virus Res.*, 105(2):121–125, October 2004.

[384] Louis-Philippe Bergeron-Sandoval, Hossein Khadivi Heris, Catherine Chang, Caitlin E Cornell, Sarah L Keller, Paul François, Adam G Hendricks, Allen J Ehrlicher, Rohit V Pappu, and Stephen W Michnick. Endocytosis caused by liquid-liquid phase separation of proteins. December 2018.

[385] Louis-Philippe Bergeron-Sandoval and Stephen W Michnick. Mechanics, structure and function of biopolymer condensates. *J. Mol. Biol.*, 430(23):4754–4761, November 2018.

[386] Erik D Holmstrom, Daniel Nettels, and Benjamin Schuler. Conformational plasticity of hepatitis C virus core protein enables RNA-Induced formation of nucleocapsid-like particles. *J. Mol. Biol.*, 430(16):2453–2467, August 2018.

[387] Lorena Rodríguez, Isabel Cuesta, Ana Asenjo, and Nieves Villanueva. Human respiratory syncytial virus matrix protein is an RNA-binding protein: binding properties, location and identity of the RNA contact residues. *J. Gen. Virol.*, 85(Pt 3):709–719, March 2004.

[388] Benjamin R Linger, Lyudmyla Kunovska, Richard J Kuhn, and Barbara L Golden. Sindbis virus nucleocapsid assembly: RNA folding promotes capsid protein dimerization. *RNA*, 10(1):128–138, January 2004.

[389] Sonia Zúñiga, Isabel Sola, Jose L Moreno, Patricia Sabella, Juan Plana-Durán, and Luis Enjuanes. Coronavirus nucleocapsid protein is an RNA chaperone. *Virology*, 357(2):215–227, January 2007.

[390] Haibin Luo, Qing Chen, Jing Chen, Kaixian Chen, Xu Shen, and Hualiang Jiang. The nucleocapsid protein of SARS coronavirus has a high binding affinity to the human cellular heterogeneous nuclear ribonucleoprotein A1. *FEBS Lett.*, 579(12):2623–2628, May 2005.

[391] Peiguo Yang, Cécile Mathieu, Regina-Maria Kolaitis, Peipei Zhang, James Messing, Ugur Yurtsever, Zemin Yang, Jinjun Wu, Yuxin Li, Qingfei Pan, Jiyang Yu, Erik W Martin, Tanja Mittag, Hong Joo Kim, and J Paul Taylor. G3BP1 is a tunable switch that triggers phase separation to assemble stress granules. *Cell*, 181(2):325–345.e28, April 2020.

[392] Mariska G M van Rosmalen, Douwe Kamsma, Andreas S Biebricher, Chenglei Li, Adam Zlotnick, Wouter H Roos, and Gijs J L Wuite. Revealing in real-time a multistep assembly mechanism for SV40 virus-like particles. *Science Advances*, 6(16):eaaz1639, April 2020.

[393] Avinash Patel, Hyun O Lee, Louise Jawerth, Shovamayee Maharana, Marcus Jahnel, Marco Y Hein, Stoyno Stoynov, Julia Mahamid, Shambaditya Saha, Titus M Franzmann, Andrej Pozniakovski, Ina Poser, Nicola Maghelli, Loic A Royer, Martin Weigert, Eugene W Myers, Stephan Grill, David Drechsel, Anthony A Hyman, and Simon Alberti. A liquid-to-solid phase transition of the ALS protein FUS accelerated by disease mutation. *Cell*, 162(5):1066–1077, August 2015.

[394] Simon Alberti and Dorothee Dormann. Liquid-Liquid phase separation in disease. *Annu. Rev. Genet.*, 53:171–194, December 2019.

[395] Stephanie C Weber and Clifford P Brangwynne. Getting RNA and protein in phase. *Cell*, 149(6):1188–1191, June 2012.

[396] X Zeng, A S Holehouse, T Mittag, A Chilkoti, and R V Pappu. Connecting coil-to-globule transitions to full phase diagrams for intrinsically disordered proteins. May 2020.

[397] Gregory L Dignon, Wenwei Zheng, Robert B Best, Young C Kim, and Jeetain Mittal. Relation between single-molecule properties and phase behavior of intrinsically disordered proteins. *Proc. Natl. Acad. Sci. U. S. A.*, 115(40):9929–9934, October 2018.

[398] Andreas Vitalis and Rohit V Pappu. ABSINTH: A new continuum solvation model for simulations of polypeptides in aqueous solutions. *J. Comput. Chem.*, 30(5):673–699, April 2009.

[399] Anuradha Mittal, Rahul K Das, Andreas Vitalis, and Rohit V Pappu. The ABSINTH implicit solvation model and forcefield paradigm for use in simulations of intrinsically disordered proteins. In Monika Fuxreiter, editor, *Computational Approaches to Protein Dynamics: From Quantum to Coarse-Grained Methods*, pages 188–222. CRC Press, 2015.

[400] Albert H Mao and Rohit V Pappu. Crystal lattice properties fully determine short-range interaction parameters for alkali and halide ions. *J. Chem. Phys.*, 137(6):064104, August 2012.

[401] Kathryn P Sherry, Rahul K Das, Rohit V Pappu, and Doug Barrick. Control of transcriptional activity by design of charge patterning in the intrinsically disordered RAM region of the notch receptor. *Proc. Natl. Acad. Sci. U. S. A.*, 114(44):E9243–E9252, October 2017.

[402] Alex S Holehouse and Rohit V Pappu. PIMMS (0.24 pre-beta), December 2019.

[403] Maxwell I. Zimmerman, Justin R. Porter, Michael D. Ward, Sukrit Singh, Neha Vithani, Artur Meller, Upasana L. Mallimadugula, Catherine E. Kuhn, Jonathan H. Borowsky, Rafal P. Wiewiora, Matthew F. D. Hurley, Aoife M Harbison, Carl A Fogarty, Joseph E.

Coffland, Elisa Fadda, Vincent A. Voelz, John D. Chodera, and Gregory R. Bowman. Citizen scientists create an exascale computer to combat covid-19. *bioRxiv*, 2020.

[404] Peng Zhou, Xing-Lou Yang, Xian-Guang Wang, Ben Hu, Lei Zhang, Wei Zhang, Hao-Rui Si, Yan Zhu, Bei Li, Chao-Lin Huang, Hui-Dong Chen, Jing Chen, Yun Luo, Hua Guo, Ren-Di Jiang, Mei-Qin Liu, Ying Chen, Xu-Rui Shen, Xi Wang, Xiao-Shuang Zheng, Kai Zhao, Quan-Jiao Chen, Fei Deng, Lin-Lin Liu, Bing Yan, Fa-Xian Zhan, Yan-Yi Wang, Geng-Fu Xiao, and Zheng-Li Shi. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*, 579(7798):270–273, March 2020.

[405] Ying Liu, Albert A Gayle, Annelies Wilder-Smith, and Joacim Rocklöv. The reproductive number of COVID-19 is higher compared to SARS coronavirus. *Journal of Travel Medicine*, 27(2):taaa021, March 2020.

[406] Gabriele Sorci, Bruno Faivre, and Serge Morand. Why does COVID-19 case fatality rate vary among countries? preprint, Infectious Diseases (except HIV/AIDS), April 2020.

[407] Morteza Abdullatif Khafaie and Fakher Rahim. Cross-country comparison of case fatality rates of covid-19/sars-cov-2. *Osong Public Health and Research Perspectives*, 11(2):74–80, April 2020.

[408] Elisabeth Mahase. Coronavirus: covid-19 has killed more people than SARS and MERS combined, despite lower case fatality rate. *BMJ*, page m641, February 2020.

[409] Graziano Onder, Giovanni Rezza, and Silvio Brusaferro. Case-fatality rate and characteristics of patients dying in relation to covid-19 in italy. *JAMA*, March 2020.

[410] Leonardo Ferreira, Ricardo dos Santos, Glaucius Oliva, and Adriano Andricopulo. Molecular docking and structure-based drug design strategies. *Molecules*, 20(7):13384–13421, July 2015.

[411] Kai J Kohlhoff, Diwakar Shukla, Morgan Lawrenz, Gregory R Bowman, David E Konerding, Dan Belov, Russ B Altman, and Vijay S Pande. Cloud-based simulations on

Google Exacycle reveal ligand modulation of GPCR activation pathways. *Nature chemistry*, 6(1):15–21, January 2014.

[412] Diwakar Shukla, Yilin Meng, Benoît Roux, and Vijay S Pande. Activation pathway of Src kinase reveals intermediate states as targets for drug design. *Nature communications*, 5(1):3397, March 2014.

[413] Kathryn M. Hart, Katelyn E. Moeder, Chris M. W. Ho, Maxwell I. Zimmerman, Thomas E. Frederick, and Gregory R. Bowman. Designing small molecules to target cryptic pockets yields both positive and negative allosteric modulators. *PLOS ONE*, 12(6):e0178678, June 2017.

[414] Wenjie Wang, Woo-Jin Shin, Bojie Zhang, Younho Choi, Ji-Seung Yoo, Maxwell I. Zimmerman, Thomas E. Frederick, Gregory R. Bowman, Michael L. Gross, Daisy W. Leung, Jae U. Jung, and Gaya K. Amarasinghe. The cap-snatching sftsv endonuclease domain is an antiviral target. *Cell Reports*, 30(1):153–163.e5, January 2020.

[415] Robert N. Kirchdoerfer, Nianshuang Wang, Jesper Pallesen, Daniel Wrapp, Hannah L. Turner, Christopher A. Cottrell, Kizzmekia S. Corbett, Barney S. Graham, Jason S. McLellan, and Andrew B. Ward. Stabilized coronavirus spikes are resistant to conformational changes induced by receptor recognition or proteolysis. *Scientific Reports*, 8(1):15701, December 2018.

[416] Alexandra C. Walls, Young-Jun Park, M. Alejandra Tortorici, Abigail Wall, Andrew T. McGuire, and David Veesler. Structure, function, and antigenicity of the sars-cov-2 spike glycoprotein. *Cell*, 181(2):281–292.e6, April 2020.

[417] Daniel Wrapp, Nianshuang Wang, Kizzmekia S. Corbett, Jory A. Goldsmith, Ching-Lin Hsieh, Olubukola Abiona, Barney S. Graham, and Jason S. McLellan. Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science*, 367(6483):1260–1263, March 2020.

[418] Haibo Zhang, Josef M. Penninger, Yimin Li, Nanshan Zhong, and Arthur S. Slutsky. Angiotensin-converting enzyme 2 (Ace2) as a SARS-CoV-2 receptor: molecular mechanisms and potential therapeutic target. *Intensive Care Medicine*, 46(4):586–590, April 2020.

[419] Markus Hoffmann, Hannah Kleine-Weber, Simon Schroeder, Nadine Krüger, Tanja Herrler, Sandra Erichsen, Tobias S. Schiergens, Georg Herrler, Nai-Huei Wu, Andreas Nitsche, Marcel A. Müller, Christian Drosten, and Stefan Pöhlmann. Sars-cov-2 cell entry depends on ace2 and tmprss2 and is blocked by a clinically proven protease inhibitor. *Cell*, 181(2):271–280.e8, April 2020.

[420] Yasunori Watanabe, Zachary T. Berndsen, Jayna Raghwani, Gemma E. Seabright, Joel D. Allen, Oliver G. Pybus, Jason S. McLellan, Ian A. Wilson, Thomas A. Bowden, Andrew B. Ward, and Max Crispin. Vulnerabilities in coronavirus glycan shields despite extensive glycosylation. *Nature Communications*, 11(1):2688, December 2020.

[421] M.I. Zimmerman and G.R. Bowman. How to run fast simulations. In *Methods in Enzymology*, volume 578, pages 213–225. Elsevier, 2016.

[422] Yuan Yuan, Duanfang Cao, Yanfang Zhang, Jun Ma, Jianxun Qi, Qihui Wang, Guangwen Lu, Ying Wu, Jinghua Yan, Yi Shi, Xinzheng Zhang, and George F. Gao. Cryo-EM structures of MERS-CoV and SARS-CoV spike glycoproteins reveal the dynamic receptor binding domains. *Nature Communications*, 8(1):15092, April 2017.

[423] Jiandong Huo, Yuguang Zhao, Jingshan Ren, Daming Zhou, Helen M.E. Duyvesteyn, Helen M. Ginn, Loic Carrique, Tomas Malinauskas, Reinis R. Ruza, Pranav N.M. Shah, Tiong Kit Tan, Pramila Rijal, Naomi Coombes, Kevin R. Bewley, Julia A. Tree, Julika Radecke, Neil G. Paterson, Piyada Supasa, Juthathip Mongkolsapaya, Gavin R. Screaton, Miles Carroll, Alain Townsend, Elizabeth E. Fry, Raymond J. Owens, and David I. Stuart. Neutralization of sars-cov-2 by destruction of the prefusion spike. *Cell Host & Microbe*, page S1931312820303516, June 2020.

[424] Ns Zhong, Bj Zheng, Ym Li, Llm Poon, Zh Xie, Kh Chan, Ph Li, Sy Tan, Q Chang, Jp Xie, Xq Liu, J Xu, Dx Li, Ky Yuen, Jsm Peiris, and Y Guan. Epidemiology and cause of severe acute respiratory syndrome (Sars) in Guangdong, People's Republic of China, in February, 2003. *The Lancet*, 362(9393):1353–1358, October 2003.

[425] Lia van der Hoek, Krzysztof Pyrc, Maarten F Jebbink, Wilma Vermeulen-Oost, Ron J M Berkhout, Katja C Wolthers, Pauline M E Wertheim-van Dillen, Jos Kaandorp, Joke Spaargaren, and Ben Berkhout. Identification of a new human coronavirus. *Nature Medicine*, 10(4):368–373, April 2004.

[426] K. Wu, W. Li, G. Peng, and F. Li. Crystal structure of NL63 respiratory coronavirus receptor-binding domain complexed with its human receptor. *Proceedings of the National Academy of Sciences*, 106(47):19970–19974, November 2009.

[427] Jian Shang, Gang Ye, Ke Shi, Yushun Wan, Chuming Luo, Hideki Aihara, Qibin Geng, Ashley Auerbach, and Fang Li. Structural basis of receptor recognition by SARS-CoV-2. *Nature*, 581(7807):221–224, May 2020.

[428] Barney S. Graham, Morgan S.A. Gilman, and Jason S. McLellan. Structure-based vaccine antigen design. *Annual Review of Medicine*, 70(1):91–104, January 2019.

[429] Linlin Zhang, Daizong Lin, Xinyuanyuan Sun, Ute Curth, Christian Drosten, Lucie Sauerhering, Stephan Becker, Katharina Rox, and Rolf Hilgenfeld. Crystal structure of SARS-CoV-2 main protease provides a basis for design of improved α-ketoamide inhibitors. *Science*, page eabb3405, March 2020.

[430] Vito Graziano, William J. McGrath, Lin Yang, and Walter F. Mangel. Sars cov main proteinase: the monomerdimer equilibrium dissociation constant. *Biochemistry*, 45(49):14632–14641, December 2006.

[431] Bhupesh Goyal and Deepti Goyal. Targeting the dimerization of the main protease of coronaviruses: a potential broad-spectrum therapeutic strategy. *ACS Combinatorial Science*, 22(6):297–305, June 2020.

[432] Dhurvas Chandrasekaran Dinesh, Dominika Chalupska, Jan Silhan, Vaclav Veverka, and Evzen Boura. Structural basis of RNA recognition by the SARS-CoV-2 nucleocapsid phosphoprotein. preprint, Biochemistry, April 2020.

[433] Jing Huang and Alexander D. MacKerell. CHARMM36 all-atom additive protein force field: Validation based on comparison to NMR data. *Journal of Computational Chemistry*, 34(25):2135–2145, September 2013.

[434] Manfred Hendlich, Friedrich Rippmann, and Gerhard Barnickel. LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins. *Journal of Molecular Graphics and Modelling*, 15(6):359–363, December 1997.

[435] The UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research*, 47(D1):D506–D515, January 2019.

[436] R. C. Edgar. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5):1792–1797, March 2004.

[437] Cameron A. Brown, Liya Hu, Zhizeng Sun, Meha P. Patel, Sukrit Singh, Justin R. Porter, Banumathi Sankaran, B. V. Venkataram Prasad, Gregory R. Bowman, and Timothy Palzkill. Antagonism between substitutions in β-lactamase explains a path not taken in the evolution of bacterial drug resistance. *Journal of Biological Chemistry*, 295(21):7376–7390, May 2020.

[438] Jeremy R. Knowles. Enzyme catalysis: not different, just better. *Nature*, 350(6314):121–124, March 1991.

[439] Michael S. Breen, Carsten Kemena, Peter K. Vlasov, Cedric Notredame, and Fyodor A. Kondrashov. Epistasis as the primary factor in molecular evolution. *Nature*, 490(7421):535–538, October 2012.

[440] J. Arjan G.M. de Visser and Joachim Krug. Empirical fitness landscapes and the predictability of evolution. *Nature Reviews Genetics*, 15(7):480–490, July 2014.

[441] James A. Wells. Additivity of mutational effects in proteins. *Biochemistry*, 29(37):8509–8517, September 1990.

[442] D. M. Weinreich. Darwinian evolution can follow only very few mutational paths to fitter proteins. *Science*, 312(5770):111–114, April 2006.

[443] Eynat Dellus-Gur, Agnes Toth-Petroczy, Mikael Elias, and Dan S. Tawfik. What makes a protein fold amenable to functional innovation? Fold polarity and stability trade-offs. *Journal of Molecular Biology*, 425(14):2609–2621, July 2013.

[444] Nicolas Doucet, Eric D. Watt, and J. Patrick Loria. The flexibility of a distant loop modulates active site motion and product release in ribonuclease a. *Biochemistry*, 48(30):7160–7168, August 2009.

[445] S. K. Whittier, A. C. Hengge, and J. P. Loria. Conformational motions regulate phosphoryl transfer in related protein tyrosine phosphatases. *Science*, 341(6148):899–903, August 2013.

[446] Leo C. James and Dan S. Tawfik. Conformational diversity and protein evolution – a 60-year-old hypothesis revisited. *Trends in Biochemical Sciences*, 28(7):361–368, July 2003.

[447] N. Tokuriki and D. S. Tawfik. Protein dynamism and evolvability. *Science*, 324(5924):203–207, April 2009.

[448] Dušan Petrović, Valeria A. Risso, Shina Caroline Lynn Kamerlin, and Jose M. Sanchez-Ruiz. Conformational dynamics and enzyme evolution. *Journal of The Royal Society Interface*, 15(144):20180330, July 2018.

[449] David M. Livermore and Neil Woodford. The β-lactamase threat in enterobacteriaceae, pseudomonas and acinetobacter. *Trends in Microbiology*, 14(9):413–420, September 2006.

[450] Jed F. Fisher, Samy O. Meroueh, and Shahriar Mobashery. Bacterial resistance to β-lactam antibiotics: compelling opportunism, compelling opportunity †. *Chemical Reviews*, 105(2):395–424, February 2005.

[451] R P Ambler, A F W Coulson, J M Frère, J M Ghuysen, B Joris, M Forsman, R C Levesque, G Tiraby, and S G Waley. A standard numbering scheme for the class A β-lactamases. *Biochemical Journal*, 276(1):269–270, May 1991.

[452] Karen Bush and Jed F. Fisher. Epidemiological expansion, structural studies, and clinical challenges of new β-lactamases from gram-negative bacteria. *Annual Review of Microbiology*, 65(1):455–478, October 2011.

[453] Timothy Palzkill. Structural and mechanistic basis for extended-spectrum drug-resistance mutations in altering the specificity of tem, ctx-m, and kpc β-lactamases. *Frontiers in Molecular Biosciences*, 5:16, February 2018.

[454] Jed F. Fisher and Shahriar Mobashery. Three decades of the class a β-lactamase acyl-enzyme, September 2009.

[455] Kinetics of β-lactamases and penicillin-binding proteins. In Bonomo and Tolmasky, editors, *Enzyme-Mediated Resistance to Antibiotics*, pages 195–213. American Society of Microbiology, January 2007.

[456] R. Bonnet. Growing group of extended-spectrum β-lactamases: the ctx-m enzymes. *Antimicrobial Agents and Chemotherapy*, 48(1):1–14, January 2004.

[457] Marco Maria D'Andrea, Fabio Arena, Lucia Pallecchi, and Gian Maria Rossolini. CTX-M-type β-lactamases: A successful story of antibiotic resistance. *International Journal of Medical Microbiology*, 303(6-7):305–317, August 2013.

[458] Yu Chen, Julien Delmas, Jacques Sirot, Brian Shoichet, and Richard Bonnet. Atomic resolution structures of ctx-m β-lactamases: extended spectrum activities from increased

mobility and decreased stability. *Journal of Molecular Biology*, 348(2):349–362, April 2005.

[459] Carolyn J. Adamski, Ana Maria Cardenas, Nicholas G. Brown, Lori B. Horton, Banumathi Sankaran, B. V. Venkataram Prasad, Hiram F. Gilbert, and Timothy Palzkill. Molecular basis for the catalytic specificity of the ctx-m extended-spectrum β-lactamases. *Biochemistry*, 54(2):447–457, January 2015.

[460] Meha P. Patel, Bartlomiej G. Fryszczyn, and Timothy Palzkill. Characterization of the global stabilizing substitution a77v and its role in the evolution of ctx-m β-lactamases. *Antimicrobial Agents and Chemotherapy*, 59(11):6741–6748, November 2015.

[461] Meha P. Patel, Liya Hu, Vlatko Stojanoski, Banumathi Sankaran, B. V. Venkataram Prasad, and Timothy Palzkill. The drug-resistant variant p167s expands the substrate profile of ctx-m β-lactamases for oxyimino-cephalosporin antibiotics by enlarging the active site upon acylation. *Biochemistry*, 56(27):3443–3453, July 2017.

[462] R. Bonnet. Effect of d240g substitution in a novel esbl ctx-m-27. *Journal of Antimicrobial Chemotherapy*, 52(1):29–35, June 2003.

[463] Soichiro Kimura, Masaji Ishiguro, Yoshikazu Ishii, Jimena Alba, and Keizo Yamaguchi. Role of a mutation at position 167 of ctx-m-19 in ceftazidime hydrolysis. *Antimicrobial Agents and Chemotherapy*, 48(5):1454–1460, May 2004.

[464] Rafael Cantón, José María González-Alba, and Juan Carlos Galán. Ctx-m enzymes: origin and diffusion. *Frontiers in Microbiology*, 3, 2012.

[465] Natalie C. J. Strynadka, Hiroyuki Adachi, Susan E. Jensen, Kathy Johns, Anita Sielecki, Christian Betzel, Kazuo Sutoh, and Michael N. G. James. Molecular structure of the acyl-enzyme intermediate in β-lactam hydrolysis at 1.7 å resolution. *Nature*, 359(6397):700–705, October 1992.

[466] Angela Novais, Rafael Canton, Teresa M. Coque, Andres Moya, Fernando Baquero, and Juan Carlos Galan. Mutational events in cefotaximase extended-spectrum β-lactamases of the ctx-m-1 cluster involved in ceftazidime resistance. *Antimicrobial Agents and Chemotherapy*, 52(7):2377–2382, July 2008.

[467] Yoshikazu Ishii, Moreno Galleni, Ling Ma, Jean-Marie Frère, and Keizo Yamaguchi. Biochemical characterisation of the CTX-M-14 β-lactamase. *International Journal of Antimicrobial Agents*, 29(2):159–164, February 2007.

[468] W. Huang and T. Palzkill. A natural polymorphism in -lactamase is a global suppressor. *Proceedings of the National Academy of Sciences*, 94(16):8801–8806, August 1997.

[469] Nicholas G. Brown, Jeanine M. Pennington, Wanzhi Huang, Tulin Ayvaz, and Timothy Palzkill. Multiple global suppressors of protein stability defects facilitate the evolution of extended-spectrum tem β-lactamases. *Journal of Molecular Biology*, 404(5):832–846, December 2010.

[470] Sebastian Mayer, Stefan Rüdiger, Hwee Ching Ang, Andreas C. Joerger, and Alan R. Fersht. Correlation of levels of folded recombinant p53 in escherichia coli with thermodynamic stability in vitro. *Journal of Molecular Biology*, 372(1):268–276, September 2007.

[471] Zheng Yuan, Timothy L. Bailey, and Rohan D. Teasdale. Prediction of protein B-factor profiles. *Proteins: Structure, Function, and Bioinformatics*, 58(4):905–912, January 2005.

[472] Meha P. Patel, Liya Hu, Cameron A. Brown, Zhizeng Sun, Carolyn J. Adamski, Vlatko Stojanoski, Banumathi Sankaran, B. V. Venkataram Prasad, and Timothy Palzkill. Synergistic effects of functionally distinct substitutions in β-lactamase variants shed light on the evolution of bacterial drug resistance. *Journal of Biological Chemistry*, 293(46):17971–17984, November 2018.

[473] Yu Chen, Richard Bonnet, and Brian K. Shoichet. The acylation mechanism of ctx-m β-lactamase at 0. 88 å resolution. *Journal of the American Chemical Society*, 129(17):5378–5380, May 2007.

[474] Monica Cartelle, Maria del Mar Tomas, Francisca Molina, Rita Moure, Rosa Villanueva, and German Bou. High-level resistance to ceftazidime conferred by a novel enzyme, ctx-m-32, derived from ctx-m-1 through a single asp240-gly substitution. *Antimicrobial Agents and Chemotherapy*, 48(6):2308–2313, June 2004.

[475] Joseph Petrosino, Carlos Cantu, and Timothy Palzkill. β-Lactamases: protein evolution in real time. *Trends in Microbiology*, 6(8):323–327, August 1998.

[476] Merijn L.M. Salverda, J. Arjan G.M. De Visser, and Miriam Barlow. Natural evolution of TEM-1 β-lactamase: experimental reconstruction and clinical relevance. *FEMS Microbiology Reviews*, 34(6):1015–1036, November 2010.

[477] Eynat Dellus-Gur, Mikael Elias, Emilia Caselli, Fabio Prati, Merijn L.M. Salverda, J. Arjan G.M. de Visser, James S. Fraser, and Dan S. Tawfik. Negative epistasis and evolvability in tem-1 β-lactamase—the thin line between an enzyme's conformational freedom and disorder. *Journal of Molecular Biology*, 427(14):2396–2409, July 2015.

[478] P. Giakkoupi. Detrimental effect of the combination of R164S with G238S in TEM-1 beta-lactamase on the extended-spectrum activity conferred by each single mutation. *Journal of Antimicrobial Chemotherapy*, 45(1):101–104, January 2000.

[479] Vlatko Stojanoski, Dar-Chone Chow, Liya Hu, Banumathi Sankaran, Hiram F. Gilbert, B. V. Venkataram Prasad, and Timothy Palzkill. A triple mutant in the -loop of tem-1 β-lactamase changes the substrate profile via a large conformational change and an altered general base for catalysis. *Journal of Biological Chemistry*, 290(16):10382–10394, April 2015.

[480] Christopher Fröhlich, Vidar Sørum, Ane Molden Thomassen, PÅl Jarle Johnsen, Hanna-Kirsti S. Leiros, and Ørjan Samuelsen. Oxa-48-mediated ceftazidime-avibactam

resistance is associated with evolutionary trade-offs. *mSphere*, 4(2):e00024–19, /msphere/4/2/mSphere024–19.atom, March 2019.

[481] Melissa D. Barnes, Magdalena A. Taracila, Joseph D. Rutter, Christopher R. Bethel, Ioannis Galdadas, Andrea M. Hujer, Emilia Caselli, Fabio Prati, John P. Dekker, Krisztina M. Papp-Wallace, Shozeb Haider, and Robert A. Bonomo. Deciphering the evolution of cephalosporin resistance to ceftolozane-tazobactam in *pseudomonas aeruginosa*. *mBio*, 9(6):e02085–18, /mbio/9/6/mBio.02085–18.atom, December 2018.

[482] Joseph Petrosino, Gary Rudgers, Hiram Gilbert, and Timothy Palzkill. Contributions of aspartate 49 and phenylalanine 142 residues of a tight binding inhibitory protein of β-lactamases. *Journal of Biological Chemistry*, 274(4):2394–2400, January 1999.

[483] Egon Amann, Jürgen Brosius, and Mark Ptashne. Vectors bearing a hybrid trp-lac promoter useful for regulated expression of cloned genes in Escherichia coli. *Gene*, 25(2-3):167–178, November 1983.

[484] F.William Studier and Barbara A. Moffatt. Use of bacteriophage T7 RNA polymerase to direct selective high-level expression of cloned genes. *Journal of Molecular Biology*, 189(1):113–130, May 1986.

[485] David C. Marciano, Jeanine M. Pennington, Xiaohu Wang, Jian Wang, Yu Chen, Veena L. Thomas, Brian K. Shoichet, and Timothy Palzkill. Genetic and structural characterization of an l201p global suppressor substitution in tem-1 β-lactamase. *Journal of Molecular Biology*, 384(1):151–164, December 2008.

[486] Paul D. Adams, Pavel V. Afonine, Gábor Bunkóczi, Vincent B. Chen, Ian W. Davis, Nathaniel Echols, Jeffrey J. Headd, Li-Wei Hung, Gary J. Kapral, Ralf W. Grosse-Kunstleve, Airlie J. McCoy, Nigel W. Moriarty, Robert Oeffner, Randy J. Read, David C. Richardson, Jane S. Richardson, Thomas C. Terwilliger, and Peter H. Zwart. *phenix* : a comprehensive python-based system for macromolecular structure solution. *Acta Crystallographica Section D Biological Crystallography*, 66(2):213–221, February 2010.

[487] P. Emsley, B. Lohkamp, W. G. Scott, and K. Cowtan. Features and development of *Coot*. *Acta Crystallographica Section D Biological Crystallography*, 66(4):486–501, April 2010.

[488] H J C Berendsen, D van der Spoel, and R van Drunen. GROMACS: A message-passing parallel molecular dynamics implementation. *Computer Physics Communications*, 91(1-3):43–56, September 1995.

[489] Alan W Sousa da Silva and Wim F Vranken. Acpype - antechamber python parser interface. *BMC Research Notes*, 5(1):367, 2012.

[490] Junmei Wang, Romain M. Wolf, James W. Caldwell, Peter A. Kollman, and David A. Case. Development and testing of a general amber force field. *Journal of Computational Chemistry*, 25(9):1157–1174, July 2004.

[491] Junmei Wang, Wei Wang, Peter A. Kollman, and David A. Case. Automatic atom type and bond type perception in molecular mechanical calculations. *Journal of Molecular Graphics and Modelling*, 25(2):247–260, October 2006.

[492] Michael D Onken, Carol M Makepeace, Kevin M Kaltenbronn, Stanley M Kanai, Tyson D Todd, Shiqi Wang, Thomas J Broekelmann, Prabakar Kumar Rao, John A Cooper, and Kendall J Blumer. Targeting nucleotide exchange to inhibit constitutively active G protein α subunits in cancer cells. *Science signaling*, 11(546):eaao6852, September 2018.

[493] Sukrit Singh and Gregory R Bowman. Quantifying allosteric communication via both concerted structural changes and conformational disorder with CARDS. *Journal of Chemical Theory and Computation*, 13(4):acs.jctc.6b01181–1517, March 2017.

[494] Ulrich Essmann, Lalith Perera, Max L Berkowitz, Tom Darden, Hsing Lee, and Lee G Pedersen. A smooth particle mesh Ewald method. *The Journal of Chemical Physics*, 103(19):8577–8593, November 1995.

[495] H J C Berendsen, J P M Postma, W F van Gunsteren, A DiNola, and J R Haak. Molecular dynamics with coupling to an external bath. *The Journal of Chemical Physics*, 81(8):3684–3690, October 1984.

[496] Alex S Holehouse and Shahar Sukenik. Controlling structural bias in intrinsically disordered proteins using solution space scanning. *J. Chem. Theory Comput.*, 16(3):1794–1805, March 2020.

[497] Zsuzsanna Dosztányi, Veronika Csizmok, Peter Tompa, and István Simon. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics*, 21(16):3433–3434, August 2005.

[498] Alex S Holehouse, Rahul K Das, James N Ahad, Mary O G Richardson, and Rohit V Pappu. CIDER: Resources to analyze Sequence-Ensemble relationships of intrinsically disordered proteins. *Biophys. J.*, 112(1):16–21, January 2017.

[499] Rahul K Das, Yongqi Huang, Aaron H Phillips, Richard W Kriwacki, and Rohit V Pappu. Cryptic sequence features within the disordered protein p27kip1 regulate cell cycle signaling. *Proc. Natl. Acad. Sci. U. S. A.*, 113(20):5616–5621, May 2016.

[500] Esther Ortega, Srinivasan Rengachari, Ziad Ibrahim, Naghmeh Hoghoughi, Jonathan Gaucher, Alex S Holehouse, Saadi Khochbin, and Daniel Panne. Transcription factor dimerization activates the p300 acetyltransferase. *Nature*, 562(7728):538–544, October 2018.

[501] Mark James Abraham, Teemu Murtola, Roland Schulz, Szilárd Páll, Jeremy C Smith, Berk Hess, and Erik Lindahl. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, 1–2:19–25, 2015.

[502] Noah S Bieler, Tuomas P J Knowles, Daan Frenkel, and Robert Vácha. Connecting macroscopic observables and microscopic assembly events in amyloid formation using coarse grained simulations. *PLoS Comput. Biol.*, 8(10):e1002692, October 2012.

[503] Steven Boeynaems, Alex S Holehouse, Venera Weinhardt, Denes Kovacs, Joris Van Lindt, Carolyn Larabell, Ludo Van Den Bosch, Rhiju Das, Peter S Tompa, Rohit V Pappu, and Aaron D Gitler. Spontaneous driving forces give rise to protein-RNA condensates with coexisting phases and complex material properties. *Proc. Natl. Acad. Sci. U. S. A.*, 116(16):7889–7898, April 2019.

[504] Kristen A Fichthorn and W H Weinberg. Theoretical foundations of dynamical monte carlo simulations. *J. Chem. Phys.*, 95(2):1090–1096, July 1991.

[505] An?ela Šarić, Alexander K Buell, Georg Meisl, Thomas C T Michaels, Christopher M Dobson, Sara Linse, Tuomas P J Knowles, and Daan Frenkel. Physical determinants of the self-replication of protein fibrils. *Nat. Phys.*, 12(9):874–880, July 2016.

[506] Benjamin Schuler, Everett A Lipman, and William A Eaton. Probing the free-energy surface for protein folding with single-molecule fluorescence spectroscopy. *Nature*, 419:743, October 2002.

[507] Hoi Sung Chung, John M Louis, and Irina V Gopich. Analysis of fluorescence lifetime and energy transfer efficiency in Single-Molecule photon trajectories of Fast-Folding proteins. *J. Phys. Chem. B*, 120(4):680–699, February 2016.

[508] Daniel Nettels, Armin Hoffmann, and Benjamin Schuler. Unfolded protein and peptide dynamics investigated with single-molecule FRET and correlation spectroscopy from picoseconds to seconds. *J. Phys. Chem. B*, 112(19):6137–6146, May 2008.

[509] Irina V Gopich, Daniel Nettels, Benjamin Schuler, and Attila Szabo. Protein dynamics from single-molecule fluorescence intensity correlation functions. *J. Chem. Phys.*, 131(9):095102, September 2009.

[510] Lothar Schäfer. *Excluded Volume Effects in Polymer Solutions: as Explained by the Renormalization Group*. Springer Science & Business Media, December 2012.

[511] Wenwei Zheng, Gül H Zerze, Alessandro Borgia, Jeetain Mittal, Benjamin Schuler, and Robert B Best. Inferring properties of disordered chains from FRET transfer efficiencies. *J. Chem. Phys.*, 148(12):123329, March 2018.

[512] Sonja Müller-Späth, Andrea Soranno, Verena Hirschfeld, Hagen Hofmann, Stefan Rüegger, Luc Reymond, Daniel Nettels, and Benjamin Schuler. Charge interactions can dominate the dimensions of intrinsically disordered proteins. *Proc. Natl. Acad. Sci. U. S. A.*, 107(33):14609–14614, August 2010.

[513] Paul G Higgs and Jean-françois Joanny. Theory of polyampholyte solutions. *J. Chem. Phys.*, 94(2):1543–1554, January 1991.

[514] Wenli Meng, Nicholas Lyle, Bowu Luan, Daniel P Raleigh, and Rohit V Pappu. Experiments and simulations show how long-range contacts can form in expanded unfolded proteins with negligible secondary structure. *Proc. Natl. Acad. Sci. U. S. A.*, 110(6):2123–2128, February 2013.

Sukrit Singh

Email: sukritsingh92@gmail.com | **Twitter**: @sukritsingh92 | **Phone**: 314.685.5659
Website: https://sukritsingh.github.io/

Education and Positions

Ph.D., Computational and Molecular Biophysics 2014 - 2020
Washington University in St. Louis
- Thesis**:** *Understanding and exploiting protein allostery & dynamics using molecular simulations*
- **Successfully defended:** August 14th, 2020
- Team member of the Folding@home consortium
- Advisor**:** Dr. Gregory R. Bowman

B.A., Chemistry and Biology 2014
Washington University in St. Louis
- Undergraduate thesis*: Synthesizing an amide bond nitroxide for improving intermolecular distance measurements.*
- Research advisor: Dr. Garland R. Marshall

Undergraduate Researcher Dec. 2012 – June 2014
Lab of Dr. Garland R. Marshall
- Development of a novel peptide mimetic of pore-forming anti-bacterials.
- Synthesized novel labels for measurement of intermolecular distances using Electron Paramagnetic Resonance.

Undergraduate Research Assistant Nov. 2010 – Nov. 2012
Lab of Dr. Joseph D. Dougherty
- Studying cerebellar apoptosis in prenatal brains as a result of glucocorticoid exposure

Teaching Experience

Teaching Assistant for General Biochemistry Aug 2015 – Dec 2015
Dept. of Biology, Washington University in St. Louis

Teaching Assistant for Modern Medicinal Chemistry Jan. 2014 – June 2014
Dept. of Chemistry, Washington University in St. Louis

Lab instructor for Introductory Organic Chemistry June 2014 – Aug. 2014
Dept. of Chemistry, Washington University in St. Louis June 2013 – Aug. 2013
- Developed new teaching experiment involving synthesis of antibacterial analogues and testing their efficacy against *E. coli*

Teaching assistant for Introduction to Computer Science Aug. 2010 – Dec 2011
Dept. of Computer Science, Washington University in St. Louis

Awards and Fellowships

Millipore-Sigma Fellowship March 2019 – 2020
WUSM Dept. of Biochemistry and Molecular Biophysics
- Provides stipend funding for 1 year and $5,000 in research funding for travel, textbooks, or equipment.

Best Poster – Runner up September 2016
Biochemistry and Molecular Biophysics Department Retreat
- Awarded for poster *Quantifying allosteric communication via structure and disorder*

MCC Travel Award September 2015
Materials Computation Center at Univ. Illinois Urbana-Champaign
- Travel funding award to attend the "Molecular and Chemical Kinetics" conference in Berlin, Germany

WU Career Center Summer Internship Award June – August 2013
Washington University in St. Louis
- Design & synthesis of novel peptidomimetics of antibacterials

HHMI Summer Undergraduate Research Fellowship June – August 2011
Washington University in St. Louis
- Investigation of glucocorticoid-induced cerebellar apoptosis

Nicolas M. Georgitsis Scholar June 2010 – May 2014
Washington University in St. Louis
- 4-year scholarship award providing full tuition and room & board

Publications

*** denotes co-first authorship**

1. Zimmerman, M.I., Porter, J.R., Ward, M.D., **Singh, S.**, Vithani, N., Meller, A., Mallimadugula, U.L, Kuhn, C. E., Borowsky, J.H., Wiewiora, R.P., Hurley, M.F.D., Harbison, A.M., Fogarty, C.A., Coffland, J.E., Fadda, E., Voelz, V.A., Chodera, J.D., Bowman, G.R. *SARS-CoV-2 Simulations Go Exascale to Capture Spike Opening and Reveal Cryptic Pockets Across the Proteome*, Accessible on BioRxiv at: https://doi.org/10.1101/2020.06.27.175430

2. Cubuk, J., Alston, J.J., Incicco, J.J., **Singh S.,** Stuchell-Brereton, M.D. , Ward, M.D., Zimmerman, M.I., Vithani, N., Griffith, D., Wagoner, J.A., Bowman, G.R., Hall, K.B., Soranno, A., Holehouse A.S., *The SARS-CoV-2 nucleocapsid protein is dynamic, disordered, and phase separates with RNA*. Submitted, Accessible on BioRxiv at: https://doi.org/10.1101/2020.06.17.158121

3. Cruz, M.A.*, Frederick, T.E.*, **Singh, S.,** Vithani, N., Zimmerman, M.I., Porter, J.R., Moeder, K.E., Amarasinghe, G.K., Bowman, G.R., *Discovery of a cryptic allosteric site in Ebola's 'undruggable' VP35 protein using simulations and experiments*, Submitted, Accessible on BioRxiv at: https://doi.org/10.1101/2020.02.09.940510

4. Brown, C.A., Hu, L., Sun, Z., Patel, M.P., **Singh, S.,** Porter, J.R., Sankaran, B., Prasad, B.V.V., Bowman, G.R., Palzkill, T.M., *Antagonism between substitutions in β-Lactamase explains a path not taken in the evolution of bacterial drug resistance.*, **Journal of Molecular Biology** (2020). doi: 10.1074/jbc.RA119.012489

5. Sun, X.* & **Singh, S.*,** Blumer, K.J., Bowman, G.R., *Simulation of spontaneous G protein activation reveals a new intermediate driving GDP unbinding*, **eLife** (2018). 7, e38465 doi: 10.7554/eLife.38465.

6. Reddy, D.N., **Singh, S.,** Ho, C.M.W., Patel, J., Schlesinger, P., Rodgers, S., Doctor, A., Marshall, G.R., *Design, synthesis, and biological evaluation of stable β6.3-Helices: Discovery of non-hemolytic antibacterial peptides.* **Eur. J. Med. Chem.** (2018). 149, 193–210

7. Patrick, G.J., Fang, L., Schaefer, J., **Singh, S.,** Bowman, G.R., Wencewicz, T.A., *Mechanistic Basis for ATP-Dependent Inhibition of Glutamine Synthetase by Tabtoxinine- β-Lactam.* **Biochemistry** (2017). 57(1), 117-135

8. **Singh S.** & Bowman, G.R., *Quantifying allosteric communication via both concerted structural changes and conformational disorder with CARDS.* **J. Chem. Theory Comput.** (2017). 13(4), 1509-1517

9. Cascella, B., Lee, S. G., **Singh, S.**, Jez, J. M. & Mirica, L. M. *The small molecule JIB-04 disrupts O2 binding in the Fe-dependent histone demethylase KDM4A/JMJD2A.* **Chem. Commun.** (2017). 53**,** 2174–2177

10. Brosey, C. A., Ho, C.M.W., Long, W.Z., **Singh**, **S.,** Burnett, K., Hura, G.L., Nix, J.C., Bowman, G.R., Ellenberger, T.E., Tainer, J.A., *Defining NADH-Driven Allostery Regulating Apoptosis-Inducing Factor*. **Structure** (2016). 24, 2067-2079

11. O'Connor, S.D., Cabrera, O.H, Dougherty, J.D., **Singh, S.**, Swiney, B.S., Salinas-Contreras, P., Farber, N.B., Noguchi, K.K, *Dexmedetomidine protects against glucocorticoid induced progenitor cell apoptosis in neonatal mouse cerebellum.* **J. Matern. Fetal. Neonatal. Med.** (2017). 30**,** 2156–2162

12. Nelson, C.A., Epperson, M.L., **Singh**, **S.**, Elliott, J.I., Fremont, D.H., *Structural Conservation and Functional Diversification within the Poxvirus Immune Evasion (PIE) Domain Superfamily.* **Viruses**, (2015). 7, 4878-4898

13. Cabrera, O.H., Dougherty, J.D., **Singh, S.,** Swiney, B.S., Farber, N.B, Noguchi, K.K, *Lithium protects against glucocorticoid induced neural progenitor cell apoptosis in the developing cerebellum.* **Brain Research**, (2014). 1545, 54-63

Invited Talks

1. **Biophysical Society 2020 Annual Meeting** — Feb. 2020
   Simulation of spontaneous G protein activation reveals a new intermediate driving GDP unbinding
   San Diego, CA., USA

2. **MilliporeSigma Fellowship lecture** — July 2019
   Allostery in cellular signaling: Capturing biological switches in action
   St. Louis, MO., USA

3. **Protein Folding Consortium 2019** — June 2019
   Identifying new intermediates in signaling proteins
   St. Louis, MO., USA

4. **Wash. U. Biochemistry and Molecular Biophysics Retreat** — October 2017
   Building an allosteric network of G protein activation via direct observation of GDP-Release
   St. Louis, MO., USA

5. **Gibbs Conference in Biothermodynamics 2016** — October 2016
   Quantifying allosteric communication via structure and disorder
   Carbondale, IL., USA

6. **Biochemistry and Molecular Biophysics Science Friday Seminar** — August 2016
   Quantifying allostery through structure and disorder: Reading the CARDS
   St. Louis, MO., USA

7. **Department of Chemistry Capstone seminar** — May 2014
   Synthesis of the Amide Bond Nitroxide and Design of Novel Heterochiral Peptide Mimetic
   St. Louis, MO., USA

8. **Midstates Consortium of Math and Sciences** — October 2013
   Synthesis of Amide Bond Nitroxide for Determination of Intermolecular Distances in HIV
   Chicago, IL., USA

Leadership & Service

Folding@home consortium — Oct. 2018 – present
Social media and outreach manager
- Manage social media (@foldingathome on twitter) and community outreach content for the Folding@home platform (https://foldingathome.org)
- Help manage collaborations with consortium partners — Sept. 2018 – present

Living Journal of Computational Molecular Sciences
Student reviewer
- Peer reviewer for articles submitted to the journal LiveComsJ (https://www.livecomsjournal.org/) — 2017 – 2018

Biochemistry and Molecular Biophysics Student Liason Committee
Chair
- Organizing seminars and events for the department — 2015 – 2019

Biochemistry and Molecular Biophysics "Science Friday" Seminar
Organizer
- Schedule speakers and host the weekly Friday department seminar

Media Appearances

Bloomberg Government — Oct. 15, 2020
Interview regarding the impact of COVID19 on academic research and careers
Link: https://about.bgov.com/news/creativity-is-simply-lost-as-covid-cripples-academic-research/

St. Louis Post Dispatch — June 26, 2020
Interview regarding Folding@home and COVID19 efforts.
Link: https://www.stltoday.com/business/local/gamers-big-tech-even-la-liga-soccer-link-computers-to-fight-covid-19-in-washington/article_ea4a4485-89f6-5140-97e1-d04e5f6e7f4a.html

Association for Computing Machinery – SIGGRAPH — May 28, 2020
Interview regarding Folding@home and COVID19 efforts.
https://blog.siggraph.org/2020/05/foldinghome-citizen-scientists-gain-insight-on-covid-19.html/

Folding@home media appearances in March while I was communications manager: https://foldingathome.org/2020/03/18/news-articles-published-in-march/ — March 2020